الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement Supérieur et de la Recherche scientifique

Université Mohamed Khider Biskra
Faculté des Sciences et de la Technologie
Département de Génie Electrique
Filière : Télécommunication
Option : Réseaux et Télécommunication

Réf:.....

Mémoire de Fin d'Etudes
En vue de l'obtention du diplôme :

MASTER

Thème

Deep Learning pour Reconnaissance du Visage

Présenté par :
TOLGUI Hocine
Soutenu le : 23 Juin 2018

Devant le jury composé de :

M[r] OUAMANE AbdElMalik	Univ. Biskra	Président
M[me] BELAHCENE Mebarka	Univ. Biskra	Encadreure
M[lle] MEDOUAKH. Saadia	Univ. Biskra	Examinatrice

Année universitaire : 2017 / 2018

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement Supérieur et de la recherche scientifique

Université Mohamed Khider Biskra
Faculté des Sciences et de la Technologie
Département de Génie Electrique
Filière : Télécommunication
Option : Réseaux et Télécommunication

Mémoire de Fin d'Etudes
En vue de l'obtention du diplôme:

MASTER

Thème

Deep Learning pour Reconnaissance du Visage

Présenté par :

TOLGUI Hocine

Avis favorable de l'encadreur :

BELAHCENE Mébarka

Avis favorable du Président du Jury

M^r^ OUAMANE Abdel Malik

Cachet et signature

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement Supérieur et de la Recherche scientifique

Université Mohamed Khider Biskra
Faculté des Sciences et de la Technologie
Département de Génie Electrique
Filière : Télécommunication
Option : Réseaux et Télécommunication

Thème :

Deep Learning pour Reconnaissance du Visage

Proposé par : BELAHCENE Mébarka
Dirigé par : BELAHCENE Mébarka

RESUMES (Français et Arabe)

Résumé

Les méthodes d'apprentissage en profondeur, en particulier les réseaux de neurones convolutifs, ont obtenu des succès significatifs dans le domaine de la vision par ordinateur. La formation de modèles profonds montre des performances exceptionnelles avec de grands ensembles de données, mais ils ne conviennent pas pour apprendre à partir de quelques échantillons limités. Ce travail de projet de fin d'étude propose un réseau de neurones à apprentissage profond pour apprendre sur un ensemble de données de taille réduite et en milieux incontrôlés. Pour cela, un ensemble de données d'image a été généré à partir du web, et il a soixante classes de personnes. Le réseau proposé est composé d'un ensemble de CNN élaborés, de RELU et de couches entièrement connectées. L'ensemble de données d'apprentissage est ensuite augmenté d'échantillons en le validant sur une BDD universelle en milieu contrôlé. Le réseau est donc formé à l'aide de la base de données standard BDD CASIA 2DV4. Nous démontrons expérimentalement que le jeu de données d'apprentissage augmenté améliore réellement la puissance de généralisation des CNN. A partir des expériences réalisées, nous avons vu que l'utilisation de cette technique était faisable, et des modifications dans l'architecture peuvent être faites comme une proposition pour améliorer la précision du modèle. En utilisant l'approche proposée pour des données d'entraînement limitées, une amélioration substantielle du taux de reconnaissance est obtenue.

ملخص

حققت أساليب التعلم المتعمق ، لا سيما الشبكات العصبية التحويلية ، نجاحا كبيرا في مجال رؤية الكمبيوتر .يُظهر تشكيل النماذج العميقة أداءً استثنائياً مع مجموعات كبيرة من البيانات ، ولكنها ليست مناسبة للتعلم من بضع عينات محدودة .يقدم مشروع نهاية الدراسة هذا شبكة من الخلايا العصبية للتعلم العميق للتعرف على مجموعة من البيانات ذات الحجم المنخفض والبيئات غير المتحكم فيها .لهذا ، تم إنشاء مجموعة من بيانات الصور من الويب ، ولديها ستين فئة من الأشخاص .تتكون الشبكة المقترحة من مجموعة من CNNs المتطورة ، RELUsوطبقات متصلة بالكامل .ثم يتم زيادة مجموعة بيانات التدريب مع العينات من خلال التحقق من صحة ذلك على BDD عالمي في بيئة يتم التحكم فيها . يتم تكوين الشبكة باستخدام قاعدة البيانات القياسية BDD 2DVنبرهن بشكل تجريبي على أن مجموعة بيانات التعلم المعززة تعمل في الواقع على تحسين قوة تعميم شبكات .CNN من التجارب التي أجريت ، رأينا أن استخدام هذا الأسلوب ممكن ، ويمكن إجراء تعديلات في الهيكل كمقترح لتحسين دقة النموذج .باستخدام النهج المقترح لبيانات التدريب المحدودة ، يتم تحقيق تحسن كبير في معدل الاعتراف.

Dédicaces

Je dédie ce mémoire à mes chers parents pour leur patience,
***Papa** ma Gratitude ne suffit pas à exprimer ce qu'elle mérite pour tous tes sacrifices depuis ma naissance, pendant mon enfance et même à l'âge adulte.*

***Ma chère mère**. Merci pour tes conseils, tes sacrifices, ton soutien et tes encouragements*

*À mes frères **Badr El Dine**, **Zakaria, Haitem**,*
*A ma Soeur **Assia***

*À tous mes amis **Rami**, **Hafid** , **Bilel** ,qui m'ont soutenu dans l'accomplissement de cet humble travail*

À tous mes professeurs et à tous ceux qui se sont engagés dans ces modestes travaux

À tout ma famille.

Hocine

Remerciements

*Tout d'abord, Louange Seigneur ''**ALLAH**'' qui nous a dotées de la merveilleuse faculté de raisonnement*

Je tiens à exprimer mes remerciements à mon encadreure

*M^me^ **BELAHCENE** Mébarka*

De m'avoir soutenu et fait confiance durant mon projet avec une grande patience. Avec son expérience dans la recherche et l'enseignement, avec ses conseils, j'ai pu découvrir le monde de la recherche scientifique dans le domaine du traitement d'image et des techniques de la biométrie faciale.

*Mes remerciements et ma profonde reconnaissance s'adressent également à M^r^ le président de jury, M^r^ **OUAMANE** AbdElMalik, d'avoir accepté de présider le jury de soutenance.*

*J'exprime également mes vifs remerciements, et ma profonde reconnaissance au Mme **MEDOUAKH** Sâadia d'avoir accepté d'examiner ce mémoire*

En second lieu, je remercie chaleureusement mes chers parents, Mon père et ma mère et mes frères, et sœur pour leurs sacrifices, aides, soutiens et encouragements et à tout ceux qui de près ou de loin ont contribués au bon déroulement de ce mémoire.

Je souhaite à présent adresser mes sincères remerciements à toutes les personnes avec qui j'ai eu la chance de travailler ou que j'ai eu l'honneur de côtoyer avant et pendant mon mémoire, et à tous les enseignants, intervenants de l'Université de BISKRA .

Liste des Figures

Chapitre 1 Système de Reconnaissance de Visage et Deep Learning

Chapitre 2 Etat de l'Art sur le Deep Learning pour la Reconnaissance

Chapitre 3 Deep Learning pour la Reconnaissance de Visage

Chapitre 4 Conception et Résultats du SRV basé sur le CNN

Liste des Tableaux

Liste des Abréviations

DL : Deep **L**earning

SVM : Support Vector Machines

DPM : modèle de pièce déformable

CNN : Convolutional Neural Network

RCNN : Region Convolutional Neural Network

ADN : Acide Désoxyribose Nucléique

FAR : False Accept Rate (Taux de fausse acceptation)

FRR : False Reject Rate (Taux de faux rejet)

EER : Error Equal Rate (Taux d'égale erreur)

RPR : Rank of Perfect Recognition

ROR : Rank One Recognition(Taux d'identification)

ROC : Receiver Operating Characteristic

CMC : Cumulative Match Characteristics

RGB (RVB) : Rouge vert bleu

TP : taux des vrais positifs

FP : taux des faux positifs

LBP : binaire local pattern

FER : reconnaissance d'expression faciale

DCNN : Deep-CNN

SRNN : réseau neuronal récurrent structurel

GNN : Gabor Neural Network

BDD : base de donnée

AI : Artificial **I**ntelligence
CV : Computer **V**ision

ANN Artificial **N**eural **N**etwork
RNN Recurrent **N**eural **N**etwork

Résumé

Les méthodes d'apprentissage en profondeur, en particulier les réseaux de neurones convolutifs, ont obtenu des succès significatifs dans le domaine de la vision par ordinateur. La formation de modèles profonds montre des performances exceptionnelles avec de grands ensembles de données, mais ils ne conviennent pas pour apprendre à partir de quelques échantillons limités. Ce travail de projet de fin d'étude propose un réseau de neurones à apprentissage profond pour apprendre sur un ensemble de données de taille réduite et en milieux incontrôlés. Pour cela, un ensemble de données d'image a été généré à partir du web, et il a soixante classes de personnes. Le réseau proposé est composé d'un ensemble de CNN élaborés, de RELU et de couches entièrement connectées. L'ensemble de données d'apprentissage est ensuite augmenté d'échantillons en le validant sur une BDD universelle en milieu contrôlé. Le réseau est donc formé à l'aide de la base de données standard BDD CASIA 2DV4. Nous démontrons expérimentalement que le jeu de données d'apprentissage augmenté améliore réellement la puissance de généralisation des CNN. A partir des expériences réalisées, nous avons vu que l'utilisation de cette technique était faisable, et des modifications dans l'architecture peuvent être faites comme une proposition pour améliorer la précision du modèle. En utilisant l'approche proposée pour des données d'entraînement limitées, une amélioration substantielle du taux de reconnaissance est obtenue.

ملخص

حققت أساليب التعلم المتعمق ، لا سيما الشبكات العصبية التحويلية ، نجاحا كبيرا في مجال رؤية الكمبيوتر .يُظهر تشكيل النماذج العميقة أداءً استثنائياً مع مجموعات كبيرة من البيانات ، ولكنها ليست مناسبة للتعلم من بضع عينات محدودة .يقدم مشروع نهاية الدراسة هذا شبكة من الخلايا العصبية للتعلم العميق للتعرف على مجموعة من البيانات ذات الحجم المنخفض والبيئات غير المتحكم فيها .لهذا ، تم إنشاء مجموعة من بيانات الصور من الويب ، ولديها ستين فئة من الأشخاص .تتكون الشبكة المقترحة من مجموعة من CNNs المتطورة ، RELUsوطبقات متصلة بالكامل .ثم يتم زيادة مجموعة بيانات التدريب مع العينات من خلال التحقق من صحة ذلك على BDD عالمي في بيئة يتم التحكم فيها .يتم تكوين الشبكة باستخدام قاعدة البيانات القياسية .BDD 2DV4 نبرهن بشكل تجريبي على أن مجموعة بيانات التعلم المعززة تعمل في الواقع على تحسين قوة تعميم شبكات .CNN من التجارب التي أجريت ، رأينا أن استخدام هذا الأسلوب ممكن ، ويمكن إجراء تعديلات في الهيكل كمقترح لتحسين دقة النموذج . باستخدام النهج المقترح لبيانات التدريب المحدودة ، يتم تحقيق تحسن كبير في معدل الاعتراف.

Introduction Générale

Table des matières

Introduction générale

Introduction Générale

1 Contexte et motivation

La reconnaissance du visage, comme un sujet de recherche interdisciplinaire, possède une longue histoire de l'étude par la communauté scientifique de la recherche depuis les années 80. Après cette période, la reconnaissance automatique de visage a parcouru un long chemin et le passage vers des produits commerciaux n'a reçu une grande impulsion qu'à partir des années 1994-1996 en grande partie grâce à la mise en œuvre d'un programme d'évaluation internationale FERET (Face Recognition Technology) sponsorisé par le ministère de la défense Américain. Généralement un système de reconnaissance de visage se compose de deux phases, la phase apprentissage (Off-line) et la phase de test (On-line). La phase d'apprentissage sera effectuée une seule fois dans laquelle l'enrôlement des images faciales des différents individus est utilisé afin d'extraire la signature biométrique de chaque individu. Ces données d'apprentissage sont préparées pour le classificateur afin de faire la reconnaissance. Au cours de la phase de test, les nouvelles données sont classifiées avec les données d'entraînement qui sont apprises dans la phase d'apprentissage. Le même traitement est effectué dans les deux phases. Pour atteindre cet objectif, généralement la procédure de traitement d'un système de reconnaissance de visage est subdivisée en trois étapes principales : détection des visages, extraction de caractéristiques, et la reconnaissance (classification).

2 Problématique

Le système de reconnaissance de visages et tous les systèmes biométriques en générale ont trois limitations principales : une limitation en termes de **performances**, une limitation en termes **d'universalité** d'utilisation et une limitation en termes de **détection des fraudes**.
L'obtention des hautes performances pour un système de reconnaissance de visage dans le monde réel est un problème ouvert pour les chercheurs depuis quelques années.

Malheureusement, les conditions non contrôlées telles que les variations d'illumination, les occultations, les expressions faciales et les variations de poses affectent considérablement les performances des systèmes de reconnaissance faciale surtout ceux qui sont basés sur l'information 2D, car ce genre d'informations dépend principalement sur les sources de la lumière, ainsi que l'image 2D ou bien l'image couleur ne représente pas la forme du visage et ne traite pas le visage comme un objet, alors que ces systèmes sont sensibles dans l'environnement réel non contrôlé. La variation de poses de la tête est un problème major pour la reconnaissance de visage. La correction de la pose et l'estimation de l'angle de rotation de la tête sont des processus nécessaires pour résoudre ce problème. Cependant, à cause de la complexité mathématique et pour des raisons de coûts élevés en termes de mémoire et de temps de calcul, le développement d'un système de reconnaissance automatique du visage robuste à la variation de poses considérées comme un grand défi pour les chercheurs de la biométrie faciale.

Les visages d'une même identité peuvent sembler très différents lorsqu'ils sont présentés dans des poses, des illuminations, des expressions, des âges et des occlusions différents. De telles variations au sein d'une même identité pourraient submerger les variations dues aux différences d'identité et rendre difficile la reconnaissance du visage, en particulier dans des conditions non contraintes. Par conséquent, la réduction des variations intra-personnelles tout en élargissant les différences inter-personnelles est un sujet central dans la reconnaissance faciale. Des études plus récentes ont également ciblé le même objectif, explicitement ou implicitement. Par exemple, l'apprentissage métrique [1,2] associe des visages à une représentation de caractéristiques telle que les faces d'une même identité sont proches les unes des autres alors que celles d'identités différentes restent séparées. Cependant, ces modèles sont très limités par leur nature linéaire ou leurs structures peu profondes, tandis que les variations inter et intra-personnelles sont complexes, hautement non linéaires et observées dans l'espace des images de haute dimension.

3 Motivation

Dans le monde de la reconnaissance faciale les ensembles de données publiques à grande échelle ont manqué et, en grande partie en raison de ce facteur, la plupart des progrès récents dans la communauté restent limités aux géants d'Internet tels que Facebook et Google (méthode de reconnaissance faciale par Google) a été formé en utilisant 200 millions d'images et huit millions d'identités uniques. La taille de cet ensemble de données est presque trois fois

plus grande que celle de tout ensemble de données de visage disponible publiquement. Les réseaux de neurones convolutionnels (CNN) ont pris d'assaut la communauté de la vision par ordinateur, améliorant considérablement l'état de l'art dans de nombreuses applications. L'un des ingrédients les plus importants pour le succès de ces méthodes est la disponibilité de grandes quantités de données d'entraînement. Le Convolutional Neural Networks (CNN) est l'une des structures réseau les plus représentatives de la technologie d'apprentissage en profondeur et a connu un grand succès dans le domaine du traitement et de la reconnaissance d'images.

4 Contribution

Au début de notre travail, beaucoup de questions et des problèmes concernant la biométrie du visage et les techniques se sont posées.

Le premier problème est de trouver une technique récente et efficace pour des systèmes monomodaux de reconnaissances de visages apte à prendre en charge des bases de données de taille considérable en milieux incontrôlés en évitant le fléau de la dimensionnalité. Pour cela nous consacrons une grande partie de notre travail (chapitre 1 et chapitre 2) pour cette recherche.

Le deuxième problème est celui des bases de données pour l'application de la technique Deep Learning dans les délais demandés.

Enfin, le problème le plus important dans notre travail est la recherche d'une méthode de reconnaissance basée sur un apprentissage automatique supervisé c'est-à-dire avec aucune connaissance à priori. Le système doit agir et évoluer d'une façon automatique pour reconnaitre les personnes. Pour toutes ces raisons, nous nous intéressons à l'outil Deep Learning particulièrement le CNN (Concolutional Neuron Network) qui assure un apprentissage en profondeur et qui pourrait répondre à toutes nos préoccupations ?

Ce document a donc pour objectif d'étudier l'architecture CNN pour l'identification et la vérification des visages. De nombreux travaux récents sur la reconnaissance faciale ont proposé de nombreuses variantes d'architectures CNN pour les visages, et nous évaluons certains de ces choix de modélisation afin de filtrer ce qui est important à partir de détails non pertinents.

Dans ce travail, nous montrons que l'apprentissage en profondeur fournit des outils beaucoup plus puissants pour gérer les deux types de variations. Grâce à son architecture profonde et à

sa grande capacité d'apprentissage, les fonctionnalités efficaces de reconnaissance faciale peuvent être apprises grâce à des mappages hiérarchiques non linéaires. Nous soutenons qu'il est essentiel d'apprendre de telles caractéristiques en utilisant simultanément deux signaux de supervision, c'est-à-dire les signaux d'identification de visage et de vérification, et les caractéristiques apprises sont appelées caractéristiques de vérification et d'identification profonde.

Notre contribution majeure est donc de se familiariser à l'outil Deep Learning et l'appliquer pour la reconnaissance de visage.

5 Structure du mémoire

Ce mémoire se présente sous forme de quatre chapitres :

Le **chapitre 1** nous présentons un aperçu sur le système de reconnaissance de visage (SRV) et du Deep Learning pour SRV.

Ensuite Le **chapitre 2** nous présentons un état de l'art récent sur l'utilisation du Deep Learning pour la vision par ordinateur et particulièrement la reconnaissance de visage.

Le **chapitre 3** est consacré à l'étude du modèle proposé suivi du **chapitre 4** concernant la conception du modèle étudié.

Nous terminons notre rédaction par une conclusion et des perspectives du présent travail.

Chapitre 1

Système de Reconnaissance de Visage et Deep Learning

Introduction

À l'origine, le mot « biométrie » peut être définie comme étant "la reconnaissance automatique d'une personne en utilisant des traits distinctifs". Une autre définition de caractéristiques biométriques est : "toutes caractéristiques physiques ou traits personnels automatiquement mesurables, robustes et distinctives qui peuvent être utilisées pour identifier un individu ou pour vérifier l'identité prétendue d'un individu" [1]. Il renvoie maintenant à un éventail de techniques, d'appareils et de systèmes permettant aux machines de reconnaître des personnes ou de confirmer ou d'authentifier leur identité [2]. Parmi toutes les technologies biométriques qui existent, la reconnaissance des visages et l'une des technologies les plus utilisées et les plus adaptées. Dans ce chapitre, nous allons mettre en relief quelques notions de base liées à la biométrie. Nous donnerons le principe de fonctionnement des systèmes biométriques, les diverses technologies et les outils utilisés pour mesurer leurs performances ainsi que les domaines d'applications. Plusieurs méthodes de reconnaissance de visages ont été proposées durant ces trente dernières années [3]. La biométrie répond aux exigences de sécurité par les secteurs particuliers et les entreprises dans tous les pays. La sécurité biométrique couvre presque tous les domaines. Aujourd'hui, la sécurité biométrique est utilisée dans l'accès aux réseaux et aux systèmes d'information, paiement électronique et cryptage des données. Généralement, les applications de la sécurité biométrique peuvent être classées en quatre sections principales :

(1) Service public : le contrôle et la sécurité des bâtiments gouvernementaux frontière ; contrôle des immigrants aux frontières ; utilisation dans les aéroports et la santé.
(2) Pouvoir judiciaire : l'utilisation des empreintes digitales pour prouver certains faits concernant les infractions pénales ; l'utilisation de l'ADN extrait du sang ou des cheveux dans la scène du crime pour obtenir le criminel.
(3) Secteurs des banques : les transactions bancaires (retraits en espèces, les cartes bancaires, paiement par le téléphone et Internet) ; la réduction de la proportion de la fraude grâce à l'intégration des cartes à puce avec reconnaissance des empreintes digitales.

Au Brésil, la police s'est préparée à la coupe du monde de football de 2014 à sa façon : elle préparait l'utilisation des lunettes équipées d'une caméra capable de filmer 400 images par seconde et les comparer avec une base de données numérique de 13 millions de photos [4]. La nouveauté dans la reconnaissance faciale arrive grâce au développement des nouvelles

caméras de type 3D. Ces caméras obtiennent de meilleurs résultats que les caméras classiques, parce qu'elles acquièrent une image tridimensionnelle de chaque visage [5]

1.1 La biométrie et les systèmes de reconnaissance de visage

Un système biométrique peut être soit un système d'identification (reconnaissance) ou un système de vérification (authentification) [6]

- ***Identification*** : consiste en une correspondance un à plusieurs (1 : N) ou bien plusieurs à plusieurs (N : N), selon le protocole de la base de données (BDD). L'utilisateur ne présente pas une identité. Le système doit trouver l'identité d'une personne parmi celles d'une base de données contenant des personnes déjà enrôlées et renvoyer l'identité correspondant à la personne se présentant devant le système, ou l'identité « inconnue » si cette personne ne fait pas partie de la base (BDD).

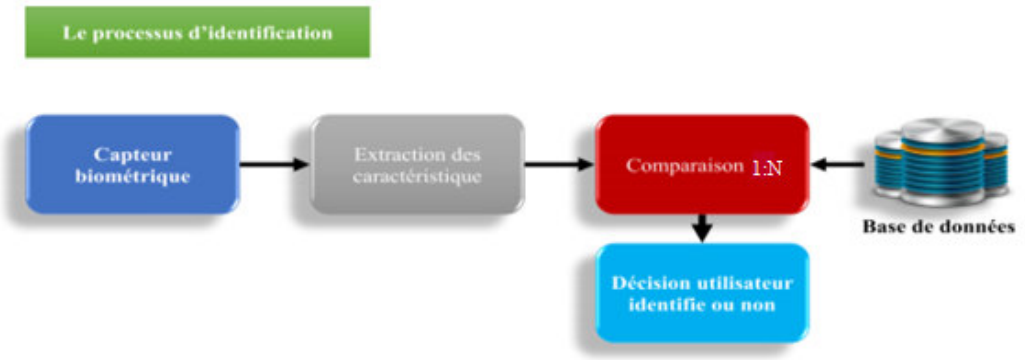


Fig.1. 1 Le mode d'identification [7]

- ***Vérification*** *(l'authentification)* : consiste en une correspondance un à un (1:1), qui compare une image du visage requête avec une image du visage modèle dont l'identité est proclamée, Le cas d'usage est une personne clamant son identité au système et celui–ci doit alors vérifier si la personne est bien la personne qu'elle prétend être. Parmi les applications liées à l'authentification, citons l'accès à des données sécurisées, des ressources informatiques ou encore des transactions sécurisées.

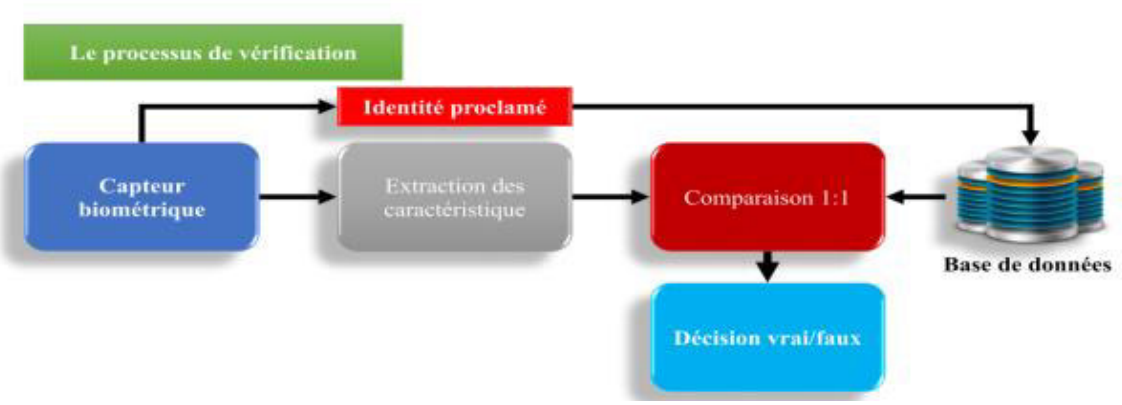


Fig.1. 2 Le mode de vérification ou authentification [7]

1.2 Architecture d'un système biométrique

Dans un système biométrique il existe toujours trois modules: le module d'apprentissage, celui de la reconnaissance et le module d'adaptation [8] [9].

1.2.1 Module d'apprentissage

Au début l'apprentissage, la caractéristique biométrique est ensuite mesurée grâce à un capteur en général, En effet, le signal contient de l'information inutile à la reconnaissance et seuls les paramètres pertinents sont extraits. Le modèle est une représentation compacte du signal qui permet de faciliter la phase de reconnaissance, mais aussi de diminuer la quantité de données à stocker.

1.2.2 Module de reconnaissance

Au cours de la reconnaissance, la caractéristique biométrique est mesurée et un ensemble de paramètres est extrait comme lors de l'apprentissage **(Fig. 1.2)**. Le capteur utilisé doit avoir des propriétés aussi proches que possibles du capteur utilisé durant la phase d'apprentissage. Il faudra en général appliquer des prétraitements supplémentaires pour limiter la dégradation des performances.

1.2.3 Module d'adaptation

Pendant la phase d'apprentissage, le système biométrique ne capture souvent que quelques instances d'un même attribut afin de limiter la gêne pour l'utilisateur, donc L'adaptation est nécessaire pour maintenir voire améliorer la performance d'un système d'utilisation.

1.3 Présentation d'un système de reconnaissance de visage

La reconnaissance automatique de visage s'effectue en trois étapes principales : ***(1) la détection de visages ; (2) l'extraction et normalisation des caractéristiques du visage et (3) l'identification et/ou vérification (reconnaissance).*** Certaines techniques de traitements d'images peuvent être communes à plusieurs étapes .Par exemple, l'extraction des caractéristiques faciales (yeux, nez, bouche) est utilisée aussi bien pour la détection que pour l'identification de visages. Par ailleurs, les étapes de détection de visage et d'extraction de caractéristiques peuvent être exécutées simultanément. Cela dépend notamment de la nature de l'application, de la taille de la base d'apprentissage, et des conditions de prise de vue (bruit, occultation, etc.). Enfin, les techniques de traitement utilisées dans chaque étape sont très critiques pour les applications biométriques, et doivent, par conséquence, être optimisées pour améliorer les performances du système global.

1.3.1 Détection de visage

La détection de visages dans l'image est un traitement indispensable et crucial avant la phase de reconnaissance. En effet, le processus de reconnaissance de visages ne pourra jamais devenir intégralement automatique s'il n'a pas été précédé par une étape de détection efficace. Un visage est considéré correctement détecté si la taille d'image extraite ne dépasse pas 20% de la taille réelle de la région faciale, et qu'elle contient essentiellement les yeux, le nez et la bouche. Elle sera ensuite affinée par un prétraitement.

1.3.2 Extraction de caractéristiques du visage

L'extraction des caractéristiques telles que les yeux, le nez, la bouche est une étape prétraitement nécessaire à la reconnaissance faciale. On peut distinguer deux pratiques différentes : la première repose sur l'extraction de régions entières du visage, elle est souvent implémentée avec une approche globale de reconnaissance de visage. La deuxième pratique extrait des points particuliers des différentes régions caractéristiques du visage, tels que les coins des yeux, de la bouche et du nez. Elle est utilisée avec une méthode locale de reconnaissance et aussi pour l'estimation de la pose du visage. Par ailleurs, plusieurs études ont été menées afin de déterminer les caractéristiques qui semblent pertinentes pour la perception, la mémorisation et la reconnaissance d'un visage humain. Les caractéristiques pertinentes rapportées sont : les cheveux, le contour du visage, les yeux et la bouche. Cette étude a également démontré le rôle important que joue le nez dans la reconnaissance faciale à partir des images de profil. En effet, dans ce cas de figure, il est évident que la forme distinctive du nez est plus intéressante que les yeux ou la bouche, les auteurs ont

particulièrement établi que la partie supérieure du visage est plus utile pour la reconnaissance faciale que la partie inférieure.

1.3.3 La reconnaissance de visage

Le module de reconnaissance exploite les caractéristiques du visage ainsi extraites pour créer une signature numérique qu'il stocke dans une base de données. Ainsi, à chaque visage de la base est associée une signature unique qui caractérise la personne correspondante. La reconnaissance d'un visage requête est obtenue par l'extraction de la signature requête correspondante et sa mise en correspondance avec la signature la plus proche dans la base de données. La reconnaissance dépend du mode de comparaison utilisé : vérification ou identification.

1.4 Système de reconnaissance de visages

Comme nous l'avons déjà cité un système automatique de reconnaissance de visages doit intégrer une étape d'apprentissage durant laquelle il associe l'allure du visage à l'identité d'une personne. Cette étape est réalisée chez les êtres humains d'une façon spontanée et évolutive. Dans un système artificiel, cette étape permet de construire une base de données des personnes connues, stockant des images étiquetées des identités. Pour ce faire, un système automatique comporte deux modes de fonctionnement : un mode enrôlement et un mode identification. Le premier mode sert à extraire pour chaque personne les éléments caractéristiques et les met sous la forme d'un vecteur caractéristique, appelé par la suite signature. Cette dernière, associée à une étiquette d'identité, sera stockée dans une base de données dédiée. Le mode d'identification permet de reconnaître une personne à partir de son image faciale, c'est à dire de retrouver l'identité associée à l'image [2].

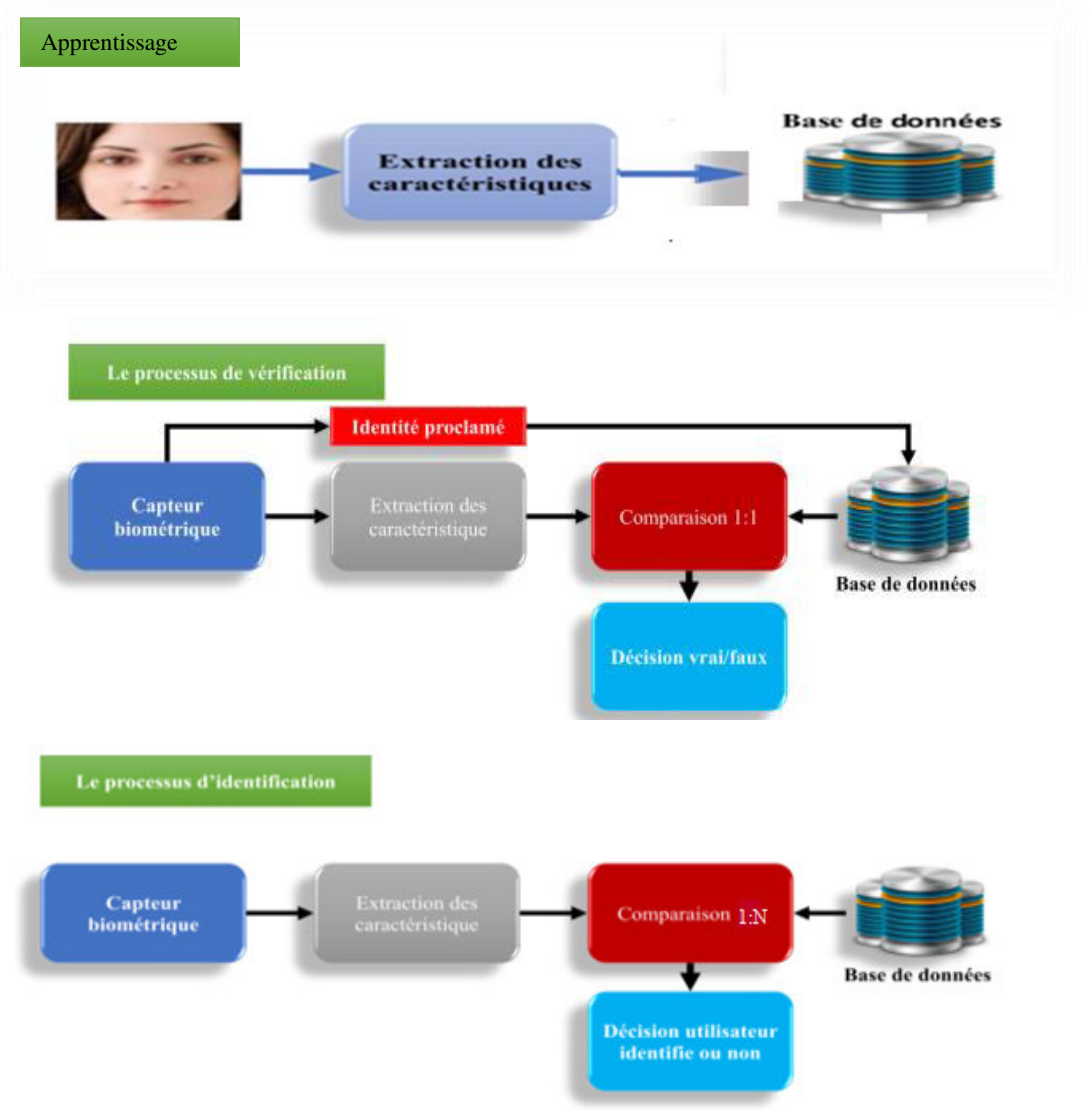


Fig.1. 3 Schéma de fonctionnement d'un système biométrique.

1.4.1 Phase d'apprentissage

La phase d'apprentissage correspondrait à un enroulement réel de personnes qui seraient enregistrées dans une base de données

Fig.1. 4 Phase d'apprentissage

1.4.2 Phase de test (Reconnaissance)

Lorsqu'une nouvelle image de la base de test est présentée, elle est comparée avec toutes les images de la phase d'apprentissage. La comparaison est faite par un calcul de la distance Euclidienne entre l'image (vecteur) de test et les images (vecteurs) de la base d'apprentissage. Il semble logique que plus la distance entre deux images est petite plus ces deux images se ressemblent. Ainsi, le résultat de la reconnaissance est l'image de la base d'apprentissage qui ressemble le plus à la nouvelle image présentée.

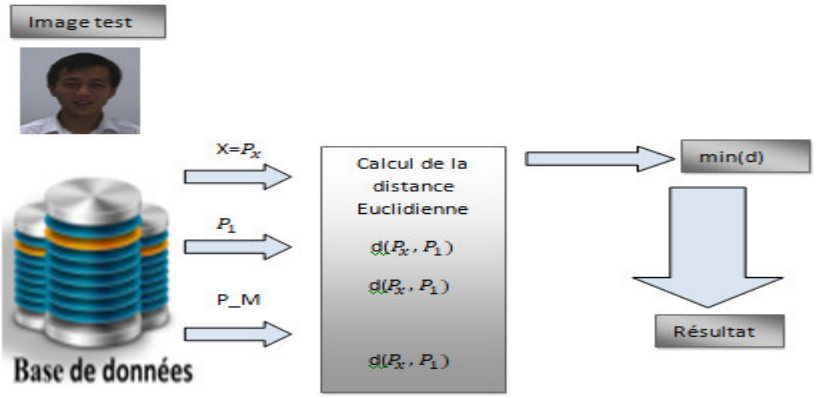


Fig.1. 5 Phase de Reconnaissance

1.5 Architecture générale

Idéalement, un système de reconnaissance faciale doit pouvoir identifier des visages présents dans une image ou une vidéo de manière automatique.
Le système peut opérer dans les deux modes suivants : authentification ou identification. On peut également noter qu'il existe un autre type de scénario de reconnaissance faciale mettant en jeu une vérification sur une liste de surveillance, où un individu est comparé à une liste restreinte de suspects.

Le principe de fonctionnement de base d'un système de reconnaissance faciale est représenté par la figure **1.3** et peut être résumé **en trois étapes** après l'acquisition. .

1.5.1 Acquisition de l'image

La **capture** est la première étape dans le processus. Il faut réussir à capter l'information pertinente sans bruit. Dans la reconnaissance de visage on peut utiliser les **capteurs 3D** par exemple pour s'affranchir des problèmes de pose. Mais leur prix excessif ne permet pas une utilisation à grande échelle. Les capteurs en proche infrarouge sont utilisés pour éliminer les problèmes de l'illumination.

Le codage consiste en l'acquisition d'image et sa digitalisation, ce qui donne lieu à une

représentation bidimensionnelle au visage. L'image dans cette étape est dans un état brut ce qui engendre un risque de bruit qui peut dégrader les performances du système et donne lieu à une représentation 2D (la matrice des niveaux de gris) [3].

1.5.2 Détection de visage et prétraitement

1.5.2.1 Détection : La détection de visage peut se faire par détection de la couleur de la peau, la forme de la tête ou par des méthodes détectant les différentes caractéristiques du visage. Cette étape est autant plus délicate quand l'image acquise contient plusieurs objets de visage ou un fond non uniforme qui crée une texture perturbant la bonne segmentation du visage. Cette étape est dépendante de la qualité des images acquises. Dans la littérature scientifique, le problème de localisation de visages est aussi désigné par la terminologie "détection de visages". Les performances globales de tout système automatique de reconnaissance dépendent amplement des performances de la détection de visages. Dans l'étape de détection, on identifie et on localise le visage dans l'image acquise au départ, indépendamment de la position, de l'échelle, de l'orientation et de l'éclairage. C'est un problème de classification où on assigne l'image à la classe visage ou à la classe non visage. On peut diviser les approches de détection en quatre catégories : les méthodes basées sur la connaissance où on code la connaissance humaine du visage, les méthodes de correspondance de masques, les méthodes à caractéristiques invariables où on utilise la couleur, les textures et les contours et finalement les méthodes les plus répandues et qui sont ceux basées sur l'apprentissage ou les statistiques comme **PCA**, **SVM**.

1.5.2.2 Prétraitement : Dans le **monde physique**, il y a trois variantes à considérer : l'éclairage, la variation de posture et l'échelle. La variation de l'un de ces trois paramètres peut conduire à une distance entre deux images du même individu, supérieure à celle séparant deux images de deux individus différents. Le rôle de cette étape est d'éliminer les parasites causés par la qualité des dispositifs optiques ou électroniques lors de l'acquisition de l'image en entrée, dans le but de ne conserver que les informations essentielles et donc préparer l'image à l'étape suivante. Elle est indispensable car on ne peut jamais avoir une image sans bruit à cause du fond (background) et de la lumière qui est généralement inconnue. Il existe plusieurs types de traitement et d'amélioration de la qualité de l'image, telle que : la **normalisation**, l'**égalisation d'histogramme**, le **filtrage**.

1.5.2.3 L'extraction des caractéristiques : Cette étape représente le cœur du système de reconnaissance, on extrait de l'image les informations qui seront sauvegardées en mémoire pour être utilisées plus tard dans la phase de décision. Le choix de ces informations utiles revient à établir un modèle pour le visage, elles doivent être discriminantes et non

redondantes. L'**analyse** est appelée indexation, représentation, modélisation ou extraction de caractéristiques. L'efficacité de cette étape a une influence directe sur la performance du système de reconnaissance de visage.

***1.5.2.4 La comparaison des caractéristiques (classification) et décision* :** Elle consiste à **modéliser** les **paramètres extraits** d'un visage ou d'un ensemble de visages d'un individu en se basant sur leurs caractéristiques communes. Un modèle est un ensemble d'informations utiles, discriminantes et non redondantes qui caractérise un ou plusieurs individus ayant des similarités, ces derniers seront regroupés dans la même classe, et ces classes varient selon le type de décision. Selon les caractéristiques extraites précédemment, les algorithmes de comparaison diffèrent.

La décision : C'est l'étape qui fait la **différence** entre un système d'**identification.** Dans cette étape, un système d'identification consiste à trouver le modèle qui correspond le mieux au visage pris en entrée à partir de ceux stockés dans la base de données, il est caractérisé par son taux de reconnaissance. Par contre, dans un système de vérification il s'agit de décider si le visage en entrée est bien celui de l'individu (modèle) **proclamé** ou il s'agit d'un **imposteur.** Pour estimer la différence entre deux images, il faut introduire une mesure de similarité

1.6 Techniques de reconnaissance de visage

Nous classons toutes ces approches en deux sous catégories à savoir : **(a)** les méthodes de sous espace (Subspace Methods) ; **(b)** les approches à base de caractéristiques géométriques (Geometric Feature Based Methods). Dans la suite, nous présentons ces deux sous catégories et les approches qui en découlent. Il est à noter que quelques-unes de ces approches ont été appliquées sur des images de profondeurs, profitant ainsi du développement mathématiques considérable que les approches 2D ont gagnées, ces quelques dernières années. Plusieurs méthodes d'identification de visages ont été proposées durant les vingt dernières années. Avant de citer les différentes techniques liées à la reconnaissance de visage **2D**, nous allons d'abord présenter un aperçu des études faites par les chercheurs en reconnaissance faciale. En effet, la connaissance des résultats de ces études est importante car elle permet le développement de nouvelles approches. Le but ultime de la reconnaissance faciale est de rivaliser, voir même dépasser, les capacités humaines de reconnaissance. Les résultats fondamentaux de ces études sont comme suit :

- Les humains peuvent reconnaître des visages familiers dans des images de faible résolution ;
- La capacité de tolérer les dégradations des images augmente avec la familiarité ;

- Les informations hautes fréquences seules, soit les contours, sont insuffisantes pour obtenir une reconnaissance faciale performante ;
- Les caractéristiques faciales sont traitées de manière holistique.

Parmi les différentes caractéristiques faciales, les sourcils sont les moins importants pour la reconnaissance ; la forme du visage est généralement codée de manière caricaturale ; la pigmentation du visage est aussi importante que sa forme ; la couleur joue un rôle important spécialement lorsque la forme est dégradée ; les changements d'illumination influencent la capacité de généralisation ; le mouvement des visages semble faciliter la reconnaissance de manière conséquente ; le système visuel progresse d'une stratégie locale vers une stratégie holistique au cours des premières années de la vie ; l'identité faciale et les expressions sont traitées par des systèmes séparés **[**10**].** On distingue trois catégories de méthodes : les ***méthodes globales***, les ***méthodes locales*** et les ***méthodes hybrides*** [11].

1.6.1 Méthodes Locales (géométriques)

Plusieurs travaux pour identifier les visages avec des traits tirés de face ont étés développés par Kanade en 1973. Les traits utilisés mesurent les différents aspects des cheveux, sourcils, yeux, bouche, etc. Le système calcule la distance entre le visage inconnu et les visages de la base, le visage ayant la plus petite distance est celui qui correspond à l'inconnu. Les résultats ont montré que seuls 6 à 7 traits sont suffisants pour identifier la plupart des visages. Parmi ces approches on peut citer : **M**odèles de **M**arkov **C**achés (**H**idden **M**arkov **M**odels (**HMM**)), l'**A**lgorithme **E**lastic **B**unch **G**raph **M**atching (**EBGM**), **E**igen Object (**EO**), l'appariement de gabarits.

1.6.2 Méthodes Globales

Cette classe regroupe les méthodes qui mettent en valeur les propriétés globales de la forme. Le visage est traité comme un tout. Parmi les approches les plus importantes réunies au sein de cette classe on trouve: L'**A**nalyse en **C**omposantes **P**rincipales (**PCA** ou **Eigen Faces**), l'**A**nalyse **D**iscriminante **L**inéaire (**LDA**), **M**achine à **V**ecteurs de **S**upport (**SVM**), les **R**éseaux de **N**eurones (**RNA**), **M**élange de **G**aussiennes (**GMM**), **M**odèle **S**urfacique du **V**isage (**3D**), l'approche **Statistique** et **Probabiliste.**

1.6.3 **Les méthodes hybrides**

Comme on a vu précédemment plusieurs approches ont été proposées pour la reconnaissance de visages, sauf qu'aucune d'elle n'est capable de s'adapter aux changements d'environnements tels que la pose, expression du visage. Pour cela, certaines approches se basent sur l'hybridation des méthodes afin d'améliorer les performances du SRV.

1.7 Exemples d'approches classiques pour la reconnaissance de visage

1.7.1 Approche ACP (visages propres)

Son but est de capturer la variation dans une collection d'images de visages et d'utiliser cette information pour coder et comparer les visages (en termes mathématique: trouver les vecteurs propres de la matrice de covariance de l'ensemble des images de visages). Le nombre possible de visages propres peut être approximé en utilisant seulement les meilleurs visages propres qui correspondent aux plus grandes valeurs propres. Cette approche rencontre le problème du coût des calculs élevé et celui de la détermination du nombre de visages propres utiles.

1.7.2 Approche Corrélation

La technique de corrélation est basée sur une comparaison simple entre une image test et les visages d'apprentissage. Celui d'entre eux se trouvant à la plus faible distance du visage test sera sélectionné comme premier choix. Malgré sa grande simplicité, cette méthode n'offre cependant pas d'avantages particulièrement intéressants. En effet, elle n'utilise pas des informations de plus haut niveau, comme la variation d'éclairage et les changements physiques.

1.7.3 Approche DCT

L'utilisation de la transformée de cosinus discrète (Discrete Cosine Transform ou DCT) à des fins de reconnaissance de visage est assez récente. Similaire aux Faces propres d'un point de vue mathématique, elle est par contre beaucoup plus rapide, tant en phase d'apprentissage qu'en phase de reconnaissance. Cela étant dit, chaque image de visage est représentée par un vecteur composé des premiers coefficients de la transformée DCT. Lorsqu'un visage est présenté au module, sa transformée est calculée et un certain nombre de coefficients est réalisée à l'aide de la distance L1 ou avec d'autres métriques pertinentes.

1.7.4 Approche Neuronal

Cette technique envisagée utilise des réseaux de neurones comme engin d'apprentissage et de reconnaissance. Pour débuter, une image brute (ou prétraitée) de dimensions fixes constitue habituellement la source d'entrés des réseaux. Les dimensions doivent être établies au préalable car le nombre de neurones sur la couche d'entrée en dépend. Cela étant dit, plus les dimensions de l'image sont élevées, plus la complexité et le temps d'apprentissage augmentent. Certains auteurs ont par ailleurs utilisé des variantes de la technique de base en modifiant les données d'entrée. Les coefficients de projections d'images dans un espace des

visages (Eigen Faces) peuvent par exemple être utilisés comme source d'information. Cette méthode peut évidemment être étendue aux coefficients de DCT, de Fourier, etc.

Tab.1. 1 Comparaison des propriétés des caractéristiques locales et globales [2]

Variations	Caractéristiques locales	Caractéristiques globales
Petites variations	Pas sensible	Sensible
Grandes variations	Sensible	Très sensible
Illuminations	Pas sensible	Sensible
Expressions	Pas sensible	Sensible
Pose	Sensible	Très sensible
Bruit	Très sensible	Sensible
Occultations	Pas sensible	Très sensible

1.8 Principales difficultés de la reconnaissance de visage

Pour le cerveau humain, le processus de la reconnaissance de visages est une tâche visuelle de haut niveau. Bien que les êtres humains puissent détecter et identifier des visages dans une scène sans beaucoup de peine, construire un système automatique qui accomplit de telles tâches représente un sérieux défi. Ce défi est d'autant plus grand lorsque les conditions d'acquisition des images sont très variables. Il existe deux types de variations associées aux images de visages : inter et intra sujet. La variation inter sujet est limitée à cause de la ressemblance physique entre les individus. Par contre la variation intra sujet est plus vaste. Elle peut être attribuée à plusieurs facteurs. Chaque visage individuel peut générer une grande variété d'images différentes. Cette grande diversité d'images de visages rend l'analyse difficile. Outre les différences générales entre les faces des variations dans l'apparence d'images de visage posent de grands problèmes à l'identification. Ces variations sont recensées comme suit:

- *Changements d'éclairage influencent l'apparition d'un visage, même si la pose de la face est fixée.*
- *Variations de pose peuvent entraîner des changements dramatiques dans les images.*
- *Les expressions faciales un outil important dans la communication humaine sont une autre source de variations dans les images. Seuls quelques points de repère du visage qui sont directement couplés avec la structure osseuse du crâne, comme la distance interoculaire ou la position générale de l'oreille sont constants dans un visage. La plupart des autres caractéristiques peuvent changer leur configuration spatiale ou*

position en raison de l'articulation de la mâchoire ou à l'action des muscles, comme les sourcils mobiles, les lèvres ou les joues.

1.8.1 Changement d'illumination

L'apparence d'un visage dans une image varie énormément en fonction de l'illumination de la scène lors de la prise de vue **(Fig. I.6)**. Les variations d'éclairage rendent la tâche de reconnaissance de visage très difficile. le changement d'apparence d'un visage dû à l'illumination, plus critique que la différence entre les individus, et peut entraîner une mauvaise classification des images d'entrée.

L'identification de visage dans un environnement non contrôlé reste donc un domaine de recherche ouvert. Les évaluations FRVT ont révélé que le problème de variation d'illumination constitue un défi majeur pour la reconnaissance faciale.

Fig.1. 6 Exemple d'un visage d'une même personne (Changement d'illumination)

[Image recueillie à partir d'Internet].

1.8.2 Variation de la pose

Dans de nombreuses applications de reconnaissance de visage, il y a toujours une différence dans l'angle de l'inclinaison de la tête entre l'image d'apprentissage et les images de tests, c.à.d. l'image d'apprentissage peut contenir une face frontale, tandis que l'image de test peut contenir un visage tourné avec un certain angle. Nous nous intéressons ici aux rotations du visage en profondeur tels que les mouvements de type hochement de tête ou négation. Le taux de reconnaissance de visage baisse considérablement quand des variations de pose sont présentes dans les images. Cette difficulté a été démontrée par des tests d'évaluation élaborés sur les bases **FERET** et **FRVT** [4].

1.8.3 Expressions faciales

L'expression faciale de l'émotion, peut produire des changements d'apparence importants des visages. Le nombre de configurations possibles est incalculable. L'influence de l'expression faciale sur la reconnaissance est donc difficile à évaluer. Puisque l'expression faciale affecte la forme géométrique et les positions des caractéristiques faciales, il semble logique que les techniques globales ou hybrides y soient plus robustes que la plupart des techniques géométriques. On soutient que les expressions faciales n'ont pas une grande influence sur les algorithmes de reconnaissance, pour autant qu'elles restent raisonnables.

Le visage d'être humain est un objet non rigide, les différentes expressions faciales comme : le sourire, le rire, la colère, la surprise et la fermeture des yeux () dégradent considérablement les performances de la reconnaissance de visage. Cela est dû au fait que l'expression faciale affecte la couleur 2D du visage et en plus les positions des éléments faciaux tels que, la bouche les yeux et les joues. Cependant, le nez est la caractéristique faciale la plus stable et la plus invariante pour ces difficultés. La déformation du visage qui est due aux expressions faciales est localisée principalement sur la partie inférieure du visage. L'information faciale se situant dans la partie supérieure du visage reste quasi invariable. Elle est généralement suffisante pour effectuer une identification. Toutefois, étant donné que l'expression faciale modifie l'aspect du visage, elle entraîne forcément une diminution du taux de reconnaissance.

Fig.1. 7 Variabilité de la présence d'expressions faciales [4]

1.8.4 Approche de reconnaissance de visage par le Deep Learning

Le problème des variantes sur les visages reste un défi et est également très difficile, et ce n'est que récemment que des résultats acceptables ont été obtenus. En fait, ce problème persiste encore surtout en environnements non contrôlés. Le système de reconnaissance de visage comme nous l'avons déjà présenté est généralement divisé en différents sous-problèmes pour faciliter le travail, principalement la détection des visages dans une image, suivie de la reconnaissance du visage.

Il existe également d'autres tâches pouvant être effectuées entre-temps, telles que les faces frontales ou l'extraction de fonctions supplémentaires. Au fil des ans, de nombreux algorithmes et techniques ont été utilisés, tels que les vecteurs propres ou les modèles Active Shape. Cependant, celui qui est actuellement le plus utilisé, et qui donne les meilleurs résultats, consiste à utiliser le Deep Learning en particulier les réseaux convolutionnels de neurones (CNN). Ces méthodes obtiennent actuellement des résultats de haute qualité, donc, après avoir passé en revue l'état actuel de l'art dans le chapitre 2, nous nous concentrons sur l'étude des réseaux de neurones convolutionnels dans le chapitre 3. L'apprentissage en profondeur est un type d'apprentissage automatique dans lequel un modèle apprend à effectuer des tâches de classification directement à partir d'images, de texte ou de son. L'apprentissage en profondeur est généralement mis en œuvre en utilisant une architecture de réseau neuronal. Le terme «profond» fait référence au nombre de couches dans le réseau : plus il y a de couches, plus le réseau est profond. Les réseaux neuronaux traditionnels ne contiennent que 2 ou 3 couches, alors que les réseaux profonds peuvent en avoir des centaines couches. On voit l'apprentissage automatique dans les programmes d'informatique, les conférences de l'industrie et le Wall Street Journal presque tous les jours. Pour tous les discours sur l'apprentissage automatique, beaucoup confondent ce qu'il peut faire avec ce qu'ils souhaitent qu'il puisse faire. Fondamentalement, l'apprentissage automatique utilise des algorithmes pour extraire des informations à partir de données brutes et les représenter dans un certain type de modèle. Nous utilisons ce modèle pour déduire des choses sur d'autres données que nous n'avons pas encore modélisées. Au milieu des années 1980 et au début des années 1990, de nombreux progrès architecturaux importants ont été réalisés dans les réseaux de neurones. Cependant, la quantité de temps et de données nécessaires pour obtenir de bons résultats a ralenti l'adoption et donc refroidi l'intérêt. Au début des années 2000, le pouvoir de calcul s'est développé de façon exponentielle et l'industrie a vu une explosion de techniques de calcul qui n'étaient pas possibles avant cela. L'apprentissage profond a émergé de la croissance informatique explosive de cette décennie comme un sérieux concurrent sur le terrain, remportant de nombreux concours d'apprentissage machine importants. Aujourd'hui, nous voyons l'apprentissage en profondeur mentionné dans tous les coins de l'apprentissage automatique, pour définir visuellement l'apprentissage en profondeur, la figure **1.8** illustre la conception de la relation entre l'IA, l'apprentissage Automatique et l'apprentissage en profondeur.

Le domaine de l'IA est vaste et existe depuis longtemps. L'apprentissage en profondeur est un sous-ensemble du domaine de l'apprentissage automatique, qui est un sous-domaine de l'IA. Jetons maintenant un coup d'œil à une autre des racines de l'apprentissage en profondeur: comment les réseaux neuronaux sont inspirés par la biologie.

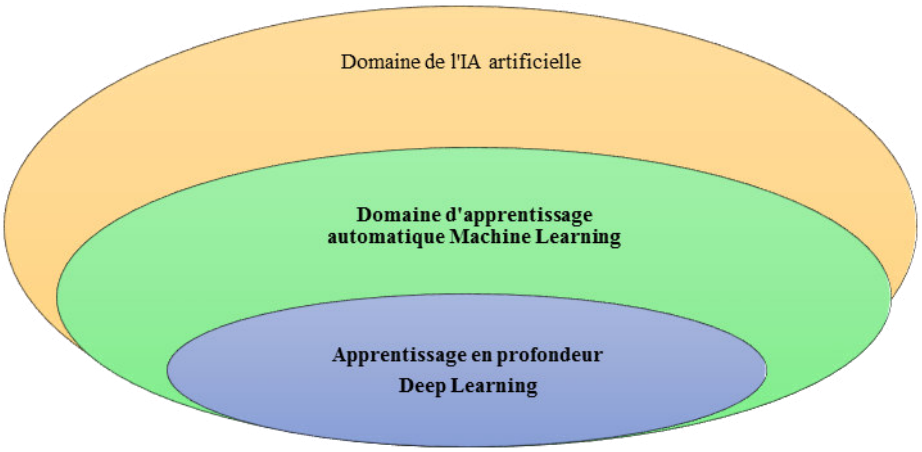


Fig.1. 8 La relation entre l'IA et l'apprentissage profond

1.9 En savoir plus sur les réseaux de neurones convolutionnels

1.9.1 À l'intérieur d'un réseau de neurones profonds

Les réseaux de neurones sont un type de modèle pour l'apprentissage automatique; ils existent depuis au moins 50 ans. L'unité fondamentale d'un réseau de neurones est un nœud, qui est vaguement basé sur le neurone biologique dans le cerveau des mammifères. Les connexions entre les neurones sont également modélisées sur des cerveaux biologiques, de même que la façon dont ces connexions se développent au cours du temps. Nous approfondirons la façon dont ces modèles fonctionnent au cours des deux prochains chapitres. Par exemple la formation d'un grand réseau de neurones convolutionnels profonds pour classer les 1,2 million d'images à haute résolution dans les 1000 classes différentes. Sur les données de test, on a atteint les taux d'erreur les plus élevés (1 et 5) de 37,5% et 17,0%, ce qui est nettement meilleur que l'état de l'art précédent. Le réseau neuronal, qui possède 60 millions de paramètres et 650 000 neurones, est constitué de cinq couches convolutionnels, dont certaines sont suivies par des couches de max-pooling, et trois couches entièrement connectées avec un softmax final de 1000 voies. Un réseau neuronal profond combine plusieurs couches de traitement non linéaires, en utilisant des éléments simples fonctionnant

en parallèle et inspirés par des systèmes biologiques. Il se compose d'une couche d'entrée, de plusieurs couches cachées et d'une couche de sortie.

Les couches sont interconnectées via des nœuds, ou neurones, chaque couche cachée utilisant la sortie de la couche précédente comme entrée, comme le figure **1.9** [12].

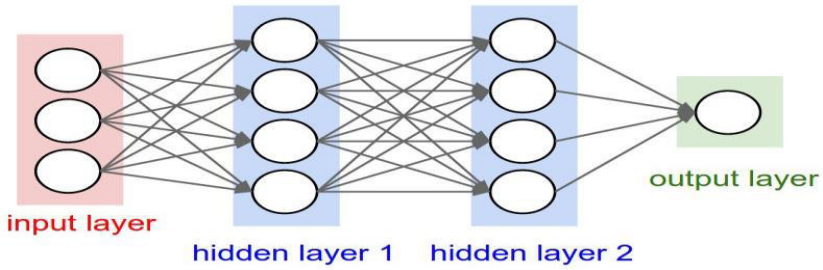


Fig.1. 9 Illustration d'un réseau de neurone

Les réseaux de neurones convolutifs (ConvNets) sont des outils largement utilisés pour l'apprentissage en profondeur. Ils sont particulièrement adaptés aux images en tant qu'entrées, bien qu'ils soient également utilisés pour d'autres applications telles que le texte, les signaux et d'autres réponses continues. Ils diffèrent des autres types de réseaux de neurones de plusieurs façons [12]. Les réseaux de neurones convolutionnels sont inspirés de la structure biologique d'un cortex visuel, qui contient des arrangements de cellules simples et complexes [13]. Ces cellules s'activent en fonction des sous-régions d'un champ visuel. Ces sous-régions sont appelées champs réceptifs. Inspirés des résultats de cette étude, les neurones dans une couche convolutive se connectent aux sous-régions des couches avant cette couche au lieu d'être entièrement connectés comme dans d'autres types de réseaux neuronaux. Les neurones ne répondent pas aux zones situées en dehors de ces sous-régions de l'image.

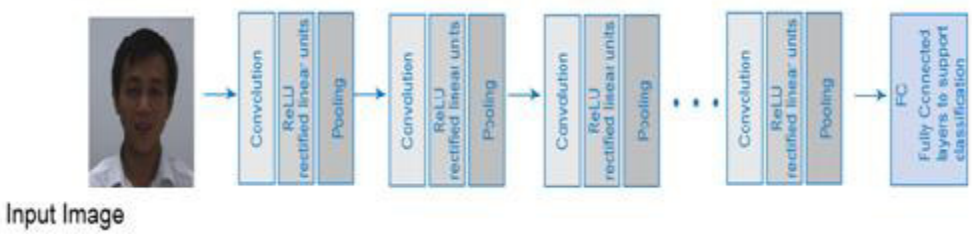


Fig.1. 10 Exemple de réseau de neurones convolutionnel CNN

Les neurones dans chaque couche d'un CNN sont disposés de manière 3D, transformant une entrée 3D en une sortie 3D. Par exemple, pour une entrée d'image, la première couche (input layer) contient les images en tant qu'entrées 3D, les dimensions étant la hauteur, la largeur et les canaux de couleur de l'image. Les neurones de la première couche convolutionnelle se connectent aux régions de ces images et les transforment en une sortie 3D. Les unités cachées (neurones) de chaque couche apprennent des combinaisons non linéaires des entrées d'origine, ce que l'on appelle l'extraction de caractéristiques [14]. Pour améliorer la performance de la reconnaissance faciale des CNN préchargés avec ou sans réglage fin, une combinaison des représentations d'images apprises par les CNN avec les fonctionnalités non CNN est effectuée. Même si les entités extraites par CNN à partir d'images de visage sont discriminantes, elles ont été désignées pour résoudre le problème de classification qui est strictement limité. Ainsi, les fonctionnalités non CNN constituent un moyen efficace pour améliorer les performances de reconnaissance des réseaux CNN. De plus, les fonctionnalités non-CNN peuvent également être utilisées avec un réglage fin pour améliorer encore les performances de la reconnaissance faciale. En menant des expériences approfondies dans [15] sur les ensembles de données LFW et CASIA3DV4.0 en utilisant le modèle VGG Face, l'outil CNN est prometteur.

1.10 Les performances des systèmes biométriques

Les paramètres suivants sont utilisés pour la mesure de la performance standard d'un système biométrique d'identification :

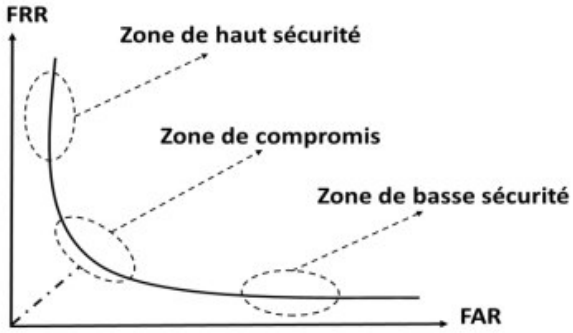
1) *Taux de reconnaissance (Rank-one Recognition Rate):* Il mesure le pourcentage des entrées qui sont correctement identifiées.

2) *Cumulative Match Characteristic (CMC):* La courbe CMC donne le pourcentage de personnes reconnues en fonction d'une variable que l'on appelle le rang [16]. On dit qu'un système reconnaît au rang 1 lorsqu'il choisit la plus proche image comme résultat de la reconnaissance. On dit qu'un système reconnaît au rang 2, lorsqu'il choisit, parmi deux images, celle qui correspond le mieux à l'image d'entrée, etc. On peut donc dire que plus le rang augmente plus le taux de reconnaissance correspondant est lié à un niveau de sécurité faible.

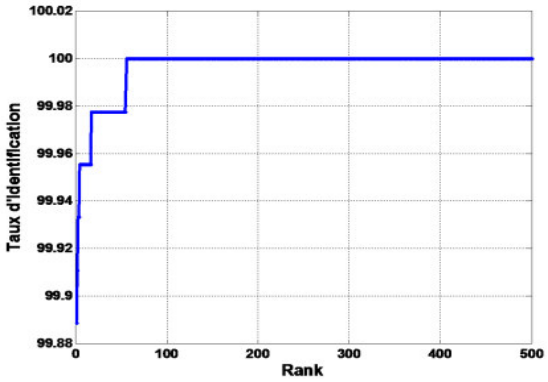
Selon le mode (*vérification* ou *identification*) du système biométrique, il existe deux façons d'un mesurer la performance :

Lorsque le système opère en mode **vérification**, on utilise ce que l'on appelle une courbe (**ROC**) "*Receiver Operating Characteristic*" représentée à la **Fig. 1.11 (a)**. Elle permet de représenter graphiquement la performance d'un système de vérification pour les différentes valeurs du seuil de décision.

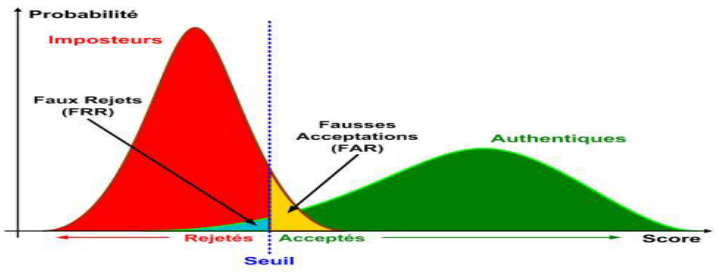
En mode **identification**, il peut être utile de savoir si le bon choix se trouve parmi les N premières réponses du système. On trace alors la *courbe* "*Cumulative Match Characteristics*" (**CMC**) qui représente la probabilité que le bon choix se trouve parmi les N premiers résultats. Comme l'illustre **la Fig. 1.11 (b)**.



(a) Courbe ROC



(b) Courbe CMC



© Illustration du FRR et du FAR

Fig.1. 11 Illustration des courbes de performances

Le taux d'erreur égal (*Equal Error Rate* ou **EER**) correspond au point **FAR = FRR**, c'est-à-dire graphiquement à l'intersection de la courbe ROC avec la première bissectrice. Il est fréquemment utilisé pour donner un aperçu de la performance d'un système. Cependant, il est important de souligner que l'EER ne résume en aucun cas toutes les caractéristiques d'un système biométrique. Le seuil θ doit donc être ajusté en fonction de l'application ciblée : haute sécurité, basse sécurité ou compromis entre les deux.

Les paramètres suivants sont utilisés pour la mesure de la performance standard d'un système biométrique avec un scénario de vérification :

1) Taux de Faux Rejet (TFR) ou (False Reject Rate, FRR): ce taux représente le pourcentage de personnes censées être reconnues mais qui sont rejetées par le système.

2) Taux de Fausse Acceptation (TFA) ou (False Accept Rate, FAR): ce taux représente le pourcentage de personnes censées ne pas être reconnues mais qui sont tout de même acceptées par le système.

3) Taux d'Égale Erreur TEE ou (Equal Error Rate, EER): Ce taux est calculé à partir des deux premiers critères et constitue un point de mesure de performance courant. Ce point correspond à l'endroit où TFR = TFA, c'est-à-dire le meilleur compromis entre les faux rejets et les fausses acceptations

TFA (F.A.R. : *False Acceptation Rate*) dans certains ouvrages, ces taux déterminent la probabilité pour un système donné de ne pas reconnaitre une personne qui normalement n'aurait pas dû être reconnue. C'est un ratio entre le nombre de personnes qui ont été acceptées alors qu'elles n'auraient pas dû l'être et le nombre total de personnes non autorisées qui ont tenté de se faire accepter.

TFR (F.R.R. : *False Reject Rate*) ces taux déterminent la probabilité pour un système donné de ne pas reconnaitre une personne qui normalement aurait être reconnue. C'est un ratio entre le nombre de personnes légitimes dont l'accès a été refusé et le nombre total de personnes légitimes s'étant présentées. La figure 1.11. (c) illustre le FRR et le FAR à partir de distributions des scores authentiques et imposteurs.

Selon la nature (authentification ou identification) du système biométrique, il existe deux façons de **mesurer la performance** en termes de Taux d'acceptation Faux **TFA** le taux de faux rejet **TFR**, défini comme suit: $$TFR = \frac{Nomber\ des\ clients\ rejetés\ (FR)}{Nomber\ total\ d'accésclient} \qquad \text{Equ. 1.1}$$

$$TFA = \frac{Nomber\ des\ imposteurs\ acceptés\ (FA)}{Nomber\ total\ d'accés\ imposteurs} \qquad \text{Equ. 1.2}$$

Une vérification parfaite d'identité ($\boldsymbol{FA = 0\ et\ FR = 0}$) est non réalisable dans la pratique. Cependant, comme montré par l'étude l'hypothèse de test binaire, n'importe lequel de ces deux taux ($\boldsymbol{TFA, TFR}$) peut être réduit à une petite valeur arbitraire en changeant le seuil de décision, avec l'inconvénient d'augmenter l'autre. Une seule mesure peut être obtenue en combinant ces deux taux d'erreurs dans le taux erreur totale ($\boldsymbol{TET}$) ou son complément, le taux de réussite total (**TR**) :

$$TR = 1 - TET \qquad \text{Equ.I. 3}$$

$$\text{TET} = \frac{\text{Nombre de Fausses Acceptation (FA)+Nombre de Faux Rejets (FR)}}{\text{Nomber total d'accés}} \qquad \text{Equ. 1.4}$$

Conclusion

Dans ce chapitre, nous avons présenté les technologies utilisées dans les systèmes biométriques pour l'identification de personnes. Cette étude nous a permis de constater que la reconnaissance de visage suscite de plus en plus l'intérêt de la communauté scientifique, car elle présente plusieurs challenges et verrous technologiques. Nous avons mis en évidence les différentes difficultés inhérentes à la reconnaissance automatique de visages, notamment l'invariance à l'illumination, pose et expressions faciales. La reconnaissance de visage en milieux incontrôlés reste un grand problème qui nécessite l'intérêt de nombreux chercheurs. Le domaine de l'intelligence artificielle et particulièrement le Deep Learning peuvent être un verrou fort intéressant pour améliorer le domaine de reconnaissance de visage car l'apprentissage profond est automatique et ne nécessite aucune phase d'apprentissage supervisé. Enfin, nous avons aussi donné un aperçu sur les techniques à base de réseaux de neurones convolutives et leur application dans les SRV (Système de Reconnaissance de Visage) ainsi que la mesure de leurs performances. Les techniques utilisées aux différentes étapes de la reconnaissance de visages sont détaillées dans les chapitres 3 et 4.

Chapitre 2

Etat de l’Art sur le Deep Learning pour la Reconnaissance

Chapitre 2 Etat de l'Art sur le Deep Learning pour la Reconnaissance

Introduction

L'intelligence artificielle ou IA pour faire court est le domaine de faire penser à l'ordinateur comme les humains en créant un cerveau artificiel. Tout ce que l'humain peut faire de façon intelligente doit être déplacé dans des machines. On peut se demander comment l'ordinateur exécute certaines tâches qui peuvent sembler difficiles à première vue, et d'autres qui peuvent sembler impossibles. Il est reconnaissable de reconnaître les visages et l'écriture ainsi que les voitures autonomes, les robots mobiles, le tri des courriels et de nombreuses autres applications [23]. Dans cet article, nous allons apprendre sur le concept de "l'apprentissage automatique" pour comprendre comment l'ordinateur peut effectuer toutes ces tâches et d'autres efficacement.

L'apprentissage automatique est classé comme une branche majeure de l'intelligence artificielle, et nous pouvons le définir comme science ce qui permet à un ordinateur de se comporter sans être programmé explicitement.

La machine fera juste ce que l'humain lui dit et pas plus. Par exemple, l'humain peut trier les nombres d'une manière intelligente et ainsi les machines devraient être intelligentes en triant les nombres comme les humains. Pour ce faire, il existe un certain nombre d'algorithmes tels que le tri à bulles qui permet à la machine de penser comme un humain. La machine suivra juste plusieurs lignes de codes qui doivent être exécutées chaque fois sans aucun changement. Il suffit de suivre les instructions que l'homme a dit à la machine sur le point de faire la tâche. La machine dans ce cas est liée à l'humain et ne peut pas fonctionner seule, Ce qui permet à un ordinateur de se comporter sans être programmé explicitement [24].

C'est comme une relation de maître et d'esclave. L'humain est le maître et la machine est l'esclave qui suit les ordres humains et pas plus. Un programme incorporant un comportement intelligent indique à la machine quoi faire. Mais l'idée d'intégrer un comportement intelligent à l'intérieur de morceaux de code ne peut pas gérer tous les comportements intelligents des humains. Certaines tâches simples comme le tri des numéros peuvent être traitées avec 100% de l'intelligence humaine. Un tel code fonctionnera avec tous les nombres même s'ils sont petits ou grands, réels ou entiers, positifs ou négatifs. Mais certaines tâches complexes ne peuvent être résolues que par du code. Il est impossible d'écrire du code pour classer des objets dans des images comme des chats, des humains, des voitures, etc. Un tel comportement

intelligent de classification des objets ne peut être simplement résolu en utilisant seulement du code car il n'y a pas de règle unique pour classer les objets. Il n'y a pas de règle pour discriminer deux classes comme les chiens et les chats en raison de l'apparence variable de ces objets et des environnements différents pour eux. Si une règle a été créée avec succès pour classer les chiens et les chats dans un environnement, elle ne peut pas fonctionner dans un autre. Mais comment rendre les machines robustes dans de telles tâches? C'est le **Machine Learning.**

Le but de l'apprentissage automatique est de programmer les ordinateurs pour qu'ils utilisent des données d'exemple ou une expérience passée pour résoudre un problème donné. De nombreuses applications réussies d'apprentissage automatique existent déjà, notamment des systèmes qui analysent les données de ventes passées pour prédire le comportement des clients, reconnaître les visages ou les paroles, optimiser le comportement du robot afin de réaliser une tâche avec des ressources minimales et extraire les connaissances bio-informatiques. Introduction au Machine Learning est un manuel complet sur le sujet, couvrant un large éventail de sujets qui ne sont généralement pas inclus dans les textes d'introduction à la machine. Il aborde de nombreuses méthodes basées sur différents domaines, notamment les statistiques, la reconnaissance de formes, les réseaux de neurones, l'intelligence artificielle, le traitement du signal, le contrôle et l'exploration de données. Tous les algorithmes d'apprentissage sont expliqués afin que l'on puisse facilement passer des équations du livre à un programme informatique. Le livre peut être utilisé par des étudiants avancés et des étudiants diplômés qui ont suivi des cours de programmation informatique, de probabilités, de calcul et d'algèbre linéaire. Après une introduction qui définit l'apprentissage automatique et donne des exemples d'applications d'apprentissage automatique, le livre couvre l'apprentissage supervisé, la théorie bayésienne de la décision, les méthodes paramétriques, méthodes multivariées, réduction de dimensionnalité, clustering, méthodes non paramétriques, arbres de décision, discrimination linéaire, perceptrons multicouches, modèles locaux, modèles de Markov cachés, évaluation et comparaison des algorithmes de classification, combinaison de plusieurs apprenants et apprentissage par renforcement.

2.1 Qu'est ce que le Machine Learning ?

Le terme **Machine Learning** (ML) fait référence à la détection automatique de motifs significatifs dans les données. Au cours des deux dernières décennies, il est devenu un outil courant dans presque toutes les tâches qui nécessitent l'extraction d'informations à partir de grands ensembles de données. Nous sommes entourés d'une technologie Machine Learning

par exemple les transactions par carte de crédit sont sécurisées pour détecter les fraudes. Les appareils photo numériques apprennent à détecter les visages et les applications intelligentes d'assistance personnelle sur les téléphones intelligents qui apprennent à reconnaître les commandes vocales. Les voitures sont équipées de systèmes de prévention des accidents construits à l'aide d'algorithmes Machine Learning.

L'apprentissage automatique (Machine Learning) est également largement utilisé dans des applications scientifiques telles que la bio-informatique, la médecine et l'astronomie.

Le premier objectif de ce chapitre est de fournir une introduction simple aux principaux concepts sous-jacents à l'apprentissage automatique: qu'est-ce que l'apprentissage? Comment une machine peut-elle apprendre? Comment quantifions-nous les ressources nécessaires pour apprendre un concept donné? L'apprentissage est-il toujours possible? Pouvons-nous savoir si le processus d'apprentissage a réussi ou échoué?

Le deuxième objectif est de présenter plusieurs algorithmes d'apprentissage machine clés. Nous avons choisi de présenter des algorithmes qui, d'une part, sont utilisés avec succès dans la pratique et, d'autre part, donnent un large éventail de techniques d'apprentissage différentes. De plus, nous prêtons une attention particulière aux algorithmes appropriés pour l'apprentissage à grande échelle, car notre monde est devenu de plus en plus numérisé ces dernières années "et la quantité de données disponibles pour l'apprentissage augmente considérablement. Le temps est le principal goulot d'étranglement. Nous quantifions donc explicitement à la fois la quantité de données et la quantité de temps de calcul nécessaire pour apprendre un concept donné.

2.1.1 Qu'est-ce que l'apprentissage ?

Quand avons-nous besoin de l'apprentissage automatique?

Quand avons-nous besoin de l'apprentissage automatique plutôt que de programmer directement nos ordinateurs pour effectuer la tâche à accomplir? Deux aspects d'un problème donné peuvent nécessiter l'utilisation de programmes qui apprennent et s'améliorent sur la base de leur expérience.

- Tâches trop complexes à programmer (tâches accomplies par les humains): Il y a de nombreuses tâches que nous, humains, accomplissons régulièrement, mais notre introspection concernant la façon dont nous les faisons n'est pas suffisamment élaborée pour extraire un programme bien défini. Des exemples de telles tâches comprennent la conduite, la reconnaissance de la parole et la compréhension de l'image. Dans toutes ces tâches, des programmes d'apprentissage automatique à la fine

pointe de la technologie, des programmes qui tirent des leçons de leur expérience, «obtiennent des résultats tout à fait satisfaisants, une fois exposés à suffisamment d'exemples de formation.

- Tâches au-delà des capacités humaines: une autre grande famille de tâches bénéficiant des techniques d'apprentissage automatique est liée à l'analyse d'ensembles de données très volumineux et complexes: données astronomiques, transformation des archives médicales en connaissances médicales, prédiction météorologique, analyse de données génomiques, moteurs de recherche Web Avec de plus en plus de données enregistrées numériquement disponibles, il devient évident qu'il existe des trésors d'informations significatives enfouies dans des archives de données qui sont beaucoup trop grandes et trop complexes pour que les humains aient du sens. Apprendre à détecter des modèles significatifs dans des ensembles de données volumineux et complexes est un domaine prometteur dans lequel la combinaison de programmes qui apprennent avec la capacité de mémoire presque illimitée et la vitesse de traitement croissante des ordinateurs ouvre de nouveaux horizons.

2.1.2 Types d'apprentissage

L'apprentissage est, bien sûr, un domaine très large. Par conséquent, le domaine de l'apprentissage automatique s'est subdivisé en plusieurs sous-domaines traitant de différents types de tâches d'apprentissage. Nous donnons une taxonomie approximative des paradigmes d'apprentissage, visant à fournir une certaine perspective de l'endroit où le contenu de ce livre se situe dans le vaste domaine de l'apprentissage automatique. Nous décrivons quatre paramètres le long desquels les paradigmes d'apprentissage peuvent être classés. Supervisé et non supervisé puisque l'apprentissage implique une interaction entre l'apprenant et l'environnement, on peut diviser les tâches d'apprentissage en fonction de la nature de cette interaction. La première distinction à noter est la différence entre l'apprentissage supervisé et non supervisé. Plus abstraitement, en considérant l'apprentissage comme un processus d'utilisation de l'expérience pour acquérir de l'expertise, l'apprentissage supervisé décrit un scénario dans lequel l'expérience, un exemple d'apprentissage, contient des informations significatives (par exemple, les étiquettes de spam).

Quelle expertise doit être appliquée ? Dans ce contexte, l'expertise acquise vise à prédire les informations manquantes pour les données de test. Dans de tels cas, nous pouvons penser à l'environnement comme un enseignant qui supervise «l'apprenant en fournissant l'information

supplémentaire (étiquettes)». Dans l'apprentissage non supervisé, cependant, il n'y a pas de distinction entre les données d'apprentissage et de test. L'apprenant traite les données d'entrée dans le but de trouver un résumé ou une version compressée de ces données. Le regroupement d'un ensemble de données en sous-ensembles d'objets similaires est un exemple typique d'une telle tâche.

2 .2 Qu'est-ce que l'apprentissage en profondeur (Deep Learning)?

L'apprentissage en profondeur est une branche de l'apprentissage automatique qui enseigne aux ordinateurs à faire ce qui vient naturellement aux humains : apprendre de l'expérience. Les algorithmes d'apprentissage automatique utilisent des méthodes de calcul pour "apprendre" des informations directement à partir de données sans s'appuyer sur une équation prédéterminée comme modèle. L'apprentissage profond est particulièrement adapté à la reconnaissance d'image, qui est importante pour résoudre des problèmes tels que la reconnaissance faciale, la détection de mouvement et de nombreuses technologies avancées d'assistance au conducteur telles que conduite autonome, détection de voie, détection de piétons et stationnement autonome [25].

Pour choisir d'utiliser un réseau pré-entraîné ou créer un nouveau réseau profond, considérons les scénarios dans ce tableau.

Tab.2. 1 *Scénarios d'utilisation ou de Deep Learning [25].*

	Utiliser un réseau prédéfini pour l'apprentissage par transfert	Créer un nouveau réseau profond
Données d'entraînement	Des centaines à des milliers d'images étiquetées (petit)	Des milliers à des millions d'images étiquetées
Calcul	Calcul modéré (GPU optionnel)	Calcul intensif (nécessite GPU pour la vitesse)
Temps de formation	Secondes à minutes	Jours à semaines pour de vrais problèmes
Précision du modèle	Bon, dépend du modèle pré-entraîné	Élevé, mais peut être adapté à de

ACCURACY		petits ensembles de données

L'apprentissage en profondeur utilise des réseaux de neurones pour apprendre des représentations utiles de caractéristiques directement à partir de données. Les réseaux de neurones combinent plusieurs couches de traitement non linéaires, en utilisant des éléments simples fonctionnant en parallèle et inspirés par des systèmes biologiques. Les modèles d'apprentissage en profondeur peuvent atteindre une précision d'avant-garde dans la classification des objets, dépassant parfois les performances au niveau humain.
On forme des modèles à l'aide d'un grand nombre de données étiquetées et d'architectures de réseau de neurones contenant de nombreuses couches, incluant généralement des couches convolutives. La formation de ces modèles est intensive en calcul et on peut généralement accélérer l'entraînement en utilisant un GPU haute performance.
De nombreuses applications d'apprentissage en profondeur utilisent des fichiers image, et parfois des millions de fichiers image. Pour accéder à de nombreux fichiers image pour un apprentissage en profondeur efficace.

L'apprentissage par transfert est couramment utilisé dans les applications d'apprentissage en profondeur. On pout utiliser un réseau pré-entraîné et l'utiliser comme point de départ pour apprendre une nouvelle tâche. La mise au point d'un réseau avec apprentissage par transfert est beaucoup plus rapide et facile que la formation à partir de rien. On peut rapidement faire en sorte que le réseau apprenne une nouvelle tâche en utilisant un plus petit nombre d'images d'entraînement. L'avantage de l'apprentissage par transfert est que le réseau pré-entraîné a déjà appris un riche ensemble de fonctionnalités qui peuvent être appliquées à un large éventail d'autres tâches similaires.
Par exemple, si l'on prend un réseau formé sur des milliers ou des millions d'images, on peut le recycler pour la détection de nouveaux objets en utilisant seulement des centaines d'images. On peut affiner efficacement un réseau pré-entraîné avec des ensembles de données beaucoup

plus petits que les données d'entraînement d'origine. Si l'on dispose d'un jeu de données très volumineux, l'apprentissage par transfert peut ne pas être plus rapide que l'apprentissage d'un nouveau réseau. L'apprentissage par transfert nous permet de :

• Transférer les caractéristiques apprises d'un réseau pré-entraîné à un nouveau problème

• Etre plus rapide et plus facile que la formation d'un nouveau réseau

• Réduire le temps de formation et la taille de l'ensemble de données

• Effectuer un apprentissage en profondeur sans avoir besoin d'apprendre à créer un nouveau réseau

L'extraction de fonctionnalités permet d'utiliser la puissance des réseaux pré-entraînés sans investir du temps et des efforts dans la formation. L'extraction de fonctionnalités peut être le moyen le plus rapide d'utiliser l'apprentissage en profondeur. On extrait des fonctions apprises d'un réseau pré-entraîné et l'on utilise ces fonctions pour former un classificateur, par exemple une machine à support de vecteurs [26].

2.2.1 Qu'est-ce qui fait de l'apprentissage en profondeur un état de l'art?

En un mot : «Accuracy ». Les outils et techniques avancés ont considérablement amélioré les algorithmes d'apprentissage en profondeur au point de surpasser les humains pour classer les images, gagner contre le meilleur joueur (GO player) du monde ou permettre à un assistant à commande vocale comme Amazon Echo et Google Home de trouver et télécharger cette nouvelle chanson qu'on aime [27].

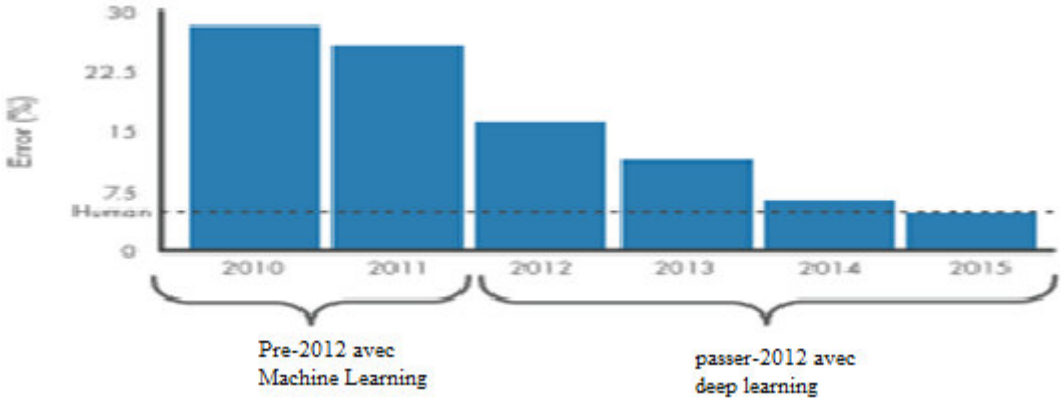


Fig.2. 1 Erreur Top-5 sur IMAGENET [27]

Trois outils technologiques rendent ce degré de précision possible :

1. ***Un accès facile à des ensembles massifs de données étiquetées :*** Des ensembles de données tels que IMAGENET sont disponibles gratuitement et sont utiles pour la formation sur de nombreux types d'objets.

2. ***Puissance de calcul accrue :*** Les GPU hautes performances accélèrent la formation des énormes quantités de données nécessaires à l'apprentissage en profondeur, réduisant ainsi le temps de formation de plusieurs semaines à quelques heures.
3. ***Modèles prédéfinis construits par des experts :*** Des modèles comme AlexNet peuvent être recyclés pour effectuer de nouvelles tâches de reconnaissance en utilisant une technique appelée apprentissage par transfert. Alors qu'AlexNet a été formé sur 1,3 million d'images haute résolution pour reconnaître 1000 objets différents, un apprentissage précis du transfert peut être réalisé avec des ensembles de données beaucoup plus petits.

2.2.2 Quelle est la différence entre apprentissage profond et apprentissage automatique?

L'apprentissage en profondeur est un sous-type de l'apprentissage automatique. Avec l'apprentissage automatique, on extrait manuellement les fonctions pertinentes d'une image. Grâce à l'apprentissage en profondeur, on alimente les images brutes directement dans un réseau neuronal profond qui apprend les fonctionnalités automatiquement.

L'apprentissage en profondeur nécessite souvent des centaines de milliers ou des millions d'images pour les meilleurs résultats. Il est également très gourmand en calcul et nécessite un GPU hautes performances.

Tab.2. 2 la différence entre apprentissage profond et apprentissage automatique [27]

Machine Learning	Depp Learning
++ De bons résultats avec de petits ensembles de données	- - nécessite de très grands ensembles de données
++ rapide pour former un modèle	- - intensément compulatoire
- - Besoin d'essayer différentes fonctionnalités et classificateurs pour obtenir les meilleurs résultats	++ apprendre les fonctionnalités et les classificateurs automatiquement
- - plateaux de précision	++ la précision est illimitée

2.3 Définitions et contexte du Deep Learning

Depuis 2006, l'apprentissage en profondeur (Deep learning), ou plus communément appelé apprentissage hiérarchique, est devenu un nouveau domaine de recherche sur l'apprentissage automatique. Au cours des dernières années, les techniques développées à partir de recherches approfondies ont déjà eu un impact sur un large éventail de travaux de traitement des signaux et de l'information dans les domaines traditionnels et nouveaux élargis incluant l'apprentissage automatique et l'intelligence artificielle (voir les articles [28, 29, 30]).

Selon **Wikipédia**, nous avons trois définitions :

Définition 1 : Une classe de techniques d'apprentissage automatique qui exploitent de nombreuses couches de traitement d'informations non linéaires pour l'extraction et la transformation de fonctions supervisées ou non supervisées, et pour l'analyse et la classification de modèles. Les caractéristiques et les concepts de niveau supérieur sont donc définis en termes de niveaux inférieurs, et une telle hiérarchie de caractéristiques est appelée architecture profonde. La plupart de ces modèles sont basés sur un apprentissage non supervisé des représentations.

Définition 2 : L'apprentissage profond fait partie d'une famille plus large de méthodes d'apprentissage automatique basées sur des représentations d'apprentissage. Une observation (par exemple, une image) peut être représentée de plusieurs façons (par exemple un vecteur de pixels), mais certaines représentations facilitent l'apprentissage des tâches, l'intérêt (par exemple, est-ce l'image d'un visage humain?), et les recherches dans ce domaine tentent de définir ce qui fait de meilleures représentations et comment les apprendre.

Définition 3 : L'apprentissage en profondeur est un ensemble d'algorithmes dans l'apprentissage automatique qui tentent d'apprendre à plusieurs niveaux, correspondant à différents niveaux d'abstraction. Il utilise généralement des réseaux de neurones artificiels. Les niveaux de ces modèles statistiques appris correspondent à des niveaux distincts de concepts, où les concepts de niveau supérieur sont définis à partir des concepts de niveau inférieur, et les mêmes concepts de niveau inférieur peuvent aider à définir de nombreux concepts de niveau supérieur [31].

Les chercheurs actifs dans ce domaine sont ceux de l'Université de Toronto, de l'Université de New York, de l'Université de Montréal, de l'Université Stanford, de Microsoft Research

(depuis 2009), Google (depuis 2011), IBM Research (depuis 2011), Facebook (depuis 2013), l'Institut de Technologie, Université de Washington, et de nombreux autres endroits [74] Voir pour une liste plus détaillée. Ces chercheurs ont démontré des succès empiriques de l'apprentissage profond dans diverses applications de vision par ordinateur, de reconnaissance phonétique, de recherche vocale, de reconnaissance faciale, de codage des caractéristiques de la parole et de l'image, de la reconnaissance manuscrite, du traitement audio et même de l'analyse des molécules pour la découverte de nouveaux médicaments comme rapporté récemment par [32] il y a un certain nombre de listes de lecture, de tutoriels, de logiciels et de conférences vidéo excellents et fréquemment mis à jour en ligne [33,34,35].

Dans ce qui suit nous présentons un aperçu de la méthodologie générale d'apprentissage en profondeur et de ses applications à diverses tâches de traitement d'image et de la reconnaissance de visage. Les domaines d'application sont choisis en fonction des trois critères suivants : (1) expertise (2) les domaines d'application qui ont déjà été transformés par l'utilisation réussie de la technologie d'apprentissage en profondeur, comme la reconnaissance faciale et la vision par ordinateur ; (3) les domaines d'application qui peuvent être significativement affectés par l'apprentissage en profondeur et qui ont connu une croissance de la recherche ; et (4) travaux récents [36].

2.4 Applications d'apprentissage en profondeur

Plusieurs études ont démontré l'efficacité des méthodes d'apprentissage en profondeur dans divers domaines d'application. Outre le défi d'écriture manuscrite MNIST [37], il existe des applications dans la détection de visages [38,39] reconnaissance et détection de la parole [40], reconnaissance générale des objets [41], traitement du langage naturel [42] et robotique. La réalité de la prolifération des données et de l'abondance de l'information sensorielle multimodale est certes un défi et un thème récurrent dans de nombreuses applications militaires et civiles, telles que les systèmes de surveillance sophistiqués. Par conséquent, l'intérêt pour l'apprentissage en profondeur n'a pas été limité à la recherche universitaire. Récemment, l'agence des projets de recherche avancée de la défense (DARPA) a annoncé un programme de recherche exclusivement axé sur l'apprentissage en profondeur.

L'apprentissage automatique profond est un domaine de recherche actif. Il reste beaucoup de travail à faire pour améliorer le processus d'apprentissage, où l'accent est actuellement mis sur l'apport d'idées fertiles provenant d'autres domaines de l'apprentissage automatique, en particulier dans le contexte de la réduction de la dimensionnalité. Un exemple comprend des

travaux récents sur le codage clairsemé [43] où la forte dimensionnalité intrinsèque des données est réduite grâce à l'utilisation de la théorie de la compression, permettant une représentation précise des signaux avec un très petit nombre de vecteurs de base.

Un autre exemple est l'apprentissage par collecteur semi-supervisé [44] où la dimensionnalité des données est réduite en mesurant la similarité entre les échantillons de données d'apprentissage, puis en projetant ces mesures de similarité dans des espaces de dimension inférieure. En outre, des approches de programmation évolutives [45,46] permettent d'approfondir l'inspiration et les techniques dans lesquelles l'apprentissage conceptuellement adaptatif et les changements architecturaux fondamentaux peuvent être appris avec un minimum d'efforts d'ingénierie.

Alors que l'apprentissage en profondeur a été appliqué avec succès à des tâches d'inférence de motifs difficiles, l'objectif du domaine dépasse de loin les applications spécifiques à une tâche. Cette portée peut rendre la comparaison de diverses méthodologies de plus en plus complexe et nécessitera probablement un effort de collaboration de la part de la communauté de recherche. Il convient également de noter que, malgré la grande perspective offerte par les technologies d'apprentissage en profondeur, certaines tâches spécifiques à un domaine peuvent ne pas être directement améliorées par de tels systèmes [48].

2.5 Travaux antérieurs : Deep Learning pour la Reconnaissance de Visage

Cette section se concentre sur la reconnaissance du visage dans les images et les vidéos, un problème qui a reçu une attention significative dans un passé récent. Parmi les nombreuses méthodes proposées dans la littérature, nous distinguons celles qui n'utilisent pas l'apprentissage profond, que nous appelons «superficielles», de celles qui le font, que nous appelons «profondes». Les méthodes peu profondes commencent par extraire une représentation de l'image du visage en utilisant des descripteurs d'image locaux artisanaux tels que SIFT, LBP, HOG [3, 6], puis ils regroupent ces descripteurs locaux en un descripteur de face global en utilisant un mécanisme de regroupement, par exemple le vecteur de Fisher [49, 50]. Il existe une grande variété de méthodes qui ne peuvent pas être décrites en détail ici (voir par exemple les références dans [49] pour une vue d'ensemble).

Notre travail concerne principalement les architectures profondes pour la reconnaissance faciale. La caractéristique déterminante de ces méthodes est l'utilisation d'un extracteur de caractéristiques CNN, une fonction apprise obtenue en composant plusieurs opérateurs linéaires et non linéaires. Un système représentatif de cette classe de méthodes est DeepFace

[51]. Cette méthode utilise un CNN profond formé pour classifier les visages en utilisant un ensemble de données de 4 millions d'exemples couvrant 4000 identités uniques. Il utilise également une architecture de réseau siamois, où le même CNN est appliqué à des paires de faces pour obtenir des descripteurs qui sont ensuite comparés en utilisant la distance euclidienne. Le but de l'entraînement est de minimiser la distance entre les paires de visages et de maximiser la distance entre les paires incongrues, une forme d'apprentissage métrique. En plus d'utiliser une très grande quantité de données d'apprentissage, DeepFace utilise un ensemble de CNN, ainsi qu'une phase de pré-traitement dans laquelle les images de visage sont alignées sur une pose canonique à l'aide d'un modèle 3D.

Lors de son introduction, DeepFace a réalisé les meilleures performances sur la référence de Labels Faces in the Wild (LFW; [30]) ainsi que sur Youtube Faces in the Wild (YFW).

Les auteurs ont ensuite étendu ce travail dans [52], en augmentant la taille de l'ensemble de données de deux ordres de grandeur, incluant 10 millions d'identités et 50 images par identité. Ils ont proposé une stratégie d'amorçage pour sélectionner les identités pour former le réseau et montré que la généralisation du réseau peut être améliorée en contrôlant la dimensionnalité de la couche entièrement connectée. Le travail de DeepFace a été prolongé par la série d'articles DeepId de Sun et al. [37, 42,53], chacun d'entre eux augmentant de façon progressive mais des performances constantes de LFW et YFW. Un certain nombre de nouvelles idées ont été incorporées dans cette série d'articles, notamment: l'utilisation de plusieurs CNN [53], un cadre d'apprentissage bayésien [54] pour former une métrique, apprentissage multitâches sur classification et vérification [45], différentes architectures CNN branchez une couche entièrement connectée après chaque couche de convolution [55], et des réseaux très profonds inspirés de [56] dans [37]. Comparé à DeepFace, Deep ID n'utilise pas l'alignement de faces 3D, mais un alignement affine 2D plus simple et s'entraîne sur la combinaison de CelebFaces [53] et WDRef [54]. Cependant, le modèle final dans [16] est assez compliqué impliquant environ 200 CNN.

2.6 Travaux récents sur le Deep Learning pour la Reconnaissance de Visage

2.6.1 Représentation dissociée :

2.6.1.1 Apprendre le GAN pour la reconnaissance de visage Pose-Invariant

Cet article [57] propose l'apprentissage de la représentation Disentangled-Generative Adversarial Network (DR-GAN) avec trois nouveautés distinctes. Premièrement, la structure codeur-décodeur du générateur permet au DR-GAN d'apprendre une représentation générative et discriminante, en plus de la synthèse d'image. Deuxièmement, cette représentation est

explicitement démêlée à d'autres variations de face telles que la pose, à travers le code de pose fourni au décodeur et l'estimation de pose dans le discriminateur. Troisièmement, DR-GAN peut prendre une ou plusieurs images comme entrée, et générer une représentation unifiée avec un nombre arbitraire d'images synthétiques.

L'évaluation quantitative et qualitative des bases de données contrôlées et non contrôlées démontre la supériorité de DR-GAN par rapport à l'état de la technique. La reconnaissance faciale est l'un des sujets les plus étudiés en vision par ordinateur [58].

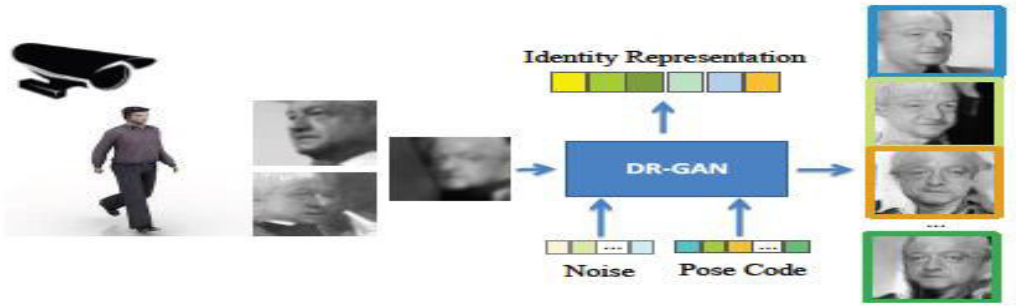


Fig.2. 2 *Avec une ou plusieurs images de visage en entrée, DR-GAN*

Comme le montre la figure 2.2, on propose l'apprentissage par représentation dissociée - réseau d'adversaires génératifs (DR-GAN) pour le PIFR (Pose Invariant Face Recognition). GAN peut générer des échantillons similaires à une distribution de données via un jeu à deux joueurs entre un générateur G et un discriminateur D. Malgré de nombreux développements prometteurs, la synthèse d'images reste l'objectif principal du GAN. Motivé par cet objectif, et le désir d'apprendre une représentation d'identité pour PIFR, les auteurs construisent G avec une structure codeur-décodeur (figure 2.3).

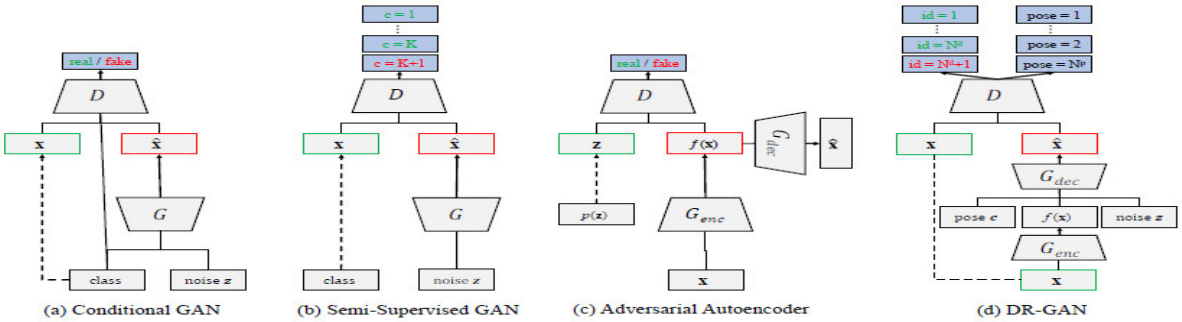


Fig.2. 3 *Comparaison des architectures GAN précédentes et de notre DR-GAN proposé [57].*

2.6.1.2 Structure du réseau

La structure de réseau de DR-GAN à image unique est présentée dans la Fig. 2.4 CASIA Net est adoptée pour G_{enc} et D.

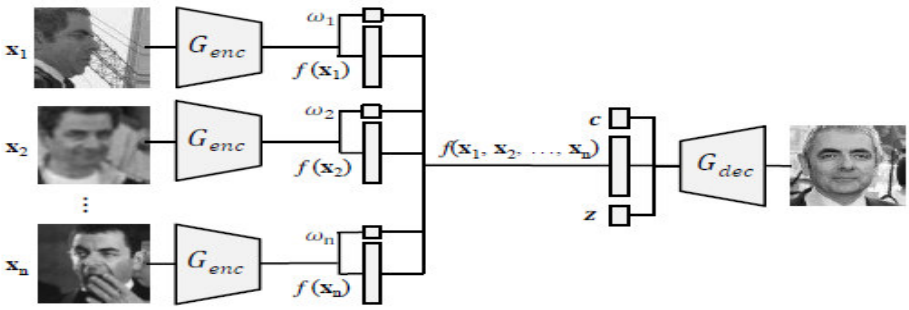


Fig.2. 4 *Générateur dans DR-GAN multi-image [57].*

À partir d'un ensemble d'images d'un sujet, nous pouvons fusionner les caractéristiques en une seule représentation via des coefficients appris dynamiquement et synthétiser des images dans n'importe quelle pose. Le but de cet article principal est la reconnaissance des visages à partir d'une seule photo ou d'un ensemble de visages suivis dans une vidéo. Ce papier a deux buts : le premier consiste à proposer une procédure permettant de créer un jeu de données de face raisonnablement grand tout en ne nécessitant qu'une quantité limitée de la personne pour l'annotation. À cette fin, une méthode de collecte de données de visage est proposée à l'aide de

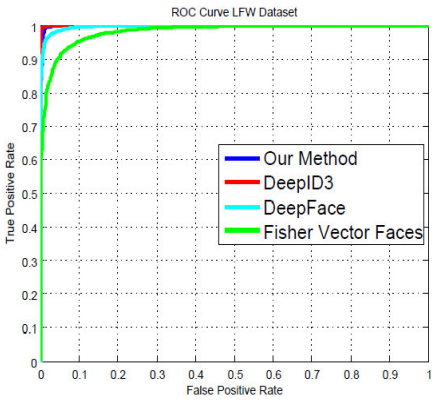
sources de connaissances disponibles sur le Web. On utilise cette procédure pour créer un ensemble de données avec plus de deux millions de faces, et le mettre gratuitement à la disposition de la communauté de la recherche, le second objectif est d'étudier diverses architectures CNN pour l'identification et la vérification des visages, y compris l'exploration de l'alignement des faces et de l'apprentissage métrique. Le tableau 2.2 compare les résultats obtenus avec les meilleurs résultats sur le jeu de données LFW, et montre également ceux-ci comme des courbes ROC. On peut observer que l'on obtient des résultats [57] comparables à l'état de l'art utilisant beaucoup moins de données et une architecture réseau beaucoup plus simple.

Tab.2. 3 *Paramètre LFW sans restriction.*

A gauche: résultats comparables à l'état de l'art nécessitant moins de données (que DeepFace et FaceNet)

A droite: courbes ROC.

No.	Method	Images	Networks	Acc.
1	Fisher Vector Faces [21]	-	-	93.10
2	DeepFace [29]	4M	3	97.35
3	Fusion [30]	500M	5	98.37
4	DeepID-2,3		200	99.47
5	FaceNet [17]	200M	1	98.87
6	FaceNet [17] + Alignment	200M	1	99.63
7	Ours	2.6M	1	98.95



Tab.2. 4 *Résultats [57] sur le jeu de données Youtube Faces.*

Résultats sur

le jeu de données Youtube Faces, paramètre non restreint

La valeur de K indique le nombre de faces utilisées pour représenter chaque vidéo.

No.	Method	Images	Networks	100%- EER	Acc.
1	Video Fisher Vector Faces [15]	-	-	87.7	83.8
2	DeepFace [29]	4M	1	91.4	91.4
3	DeepID-2,2+,3		200	-	93.2
4	FaceNet [17] + Alignment	200M	1	-	95.1
5	Ours ($K = 100$)	2.6M	1	92.8	91.6
6	Ours ($K = 100$) + Embedding learning	2.6M	1	97.4	97.3

Notons que les résultats de DeepID3 sont pour l'ensemble de test avec des erreurs d'étiquette corrigées qui n'ont pas été effectuées par une autre méthode. Dans ce travail, deux contributions ont été réalisées fait: premièrement, une conception de la procédure qui est capable d'assembler un ensemble de données à grande échelle, avec un faible bruit d'étiquette, tout en minimisant la quantité d'annotation manuelle impliquée. L'une des idées clés consistait à utiliser des classificateurs plus faibles pour classer les données présentées aux annotateurs. Cette procédure a été développée pour les visages, mais est évidemment adaptée à d'autres classes d'objets ainsi qu'à des tâches à granularité fine. La deuxième contribution a été de montrer qu'un CNN profond, sans aucun embellissement mais avec une formation appropriée, peut atteindre des résultats comparables à l'état de la technique. Encore une fois, c'est une conclusion qui peut être applicable à de nombreuses autres tâches.

2.6.2 Reconnaissance de visage basée sur le réseau de neurones convolutif

La structure générale du processus de reconnaissance faciale dans cet article [59] est composée de trois étapes. Il commence par l'étape de prétraitement: conversion de l'espace colorimétrique et redimensionnement des images, poursuite de l'extraction des traits faciaux, puis classification des éléments extraits. Dans ce système, Softmax Classifier doit réaliser l'étape finale qui est la classification sur la base des traits faciaux extraits de CNN.

2.6.6.1 Méthodologie : Les CNN sont une catégorie de réseaux de neurones qui se sont révélés très efficaces dans des domaines tels que la reconnaissance d'image et la classification. Les réseaux CNN sont un type de réseaux neuronaux d'anticipation constitués de plusieurs couches. Les CNN sont constitués de filtres ou de noyaux ou de neurones qui ont des poids ou des paramètres apprenants et des biais. Chaque filtre prend des entrées, effectue une convolution et le suit éventuellement avec une non-linéarité [60]. Une architecture CNN typique peut être vue comme le montre la Fig. 2.5 La structure de CNN contient des couches Convolutional, Pooling, Rectified Linear Unit (ReLU) et Fully Connected.

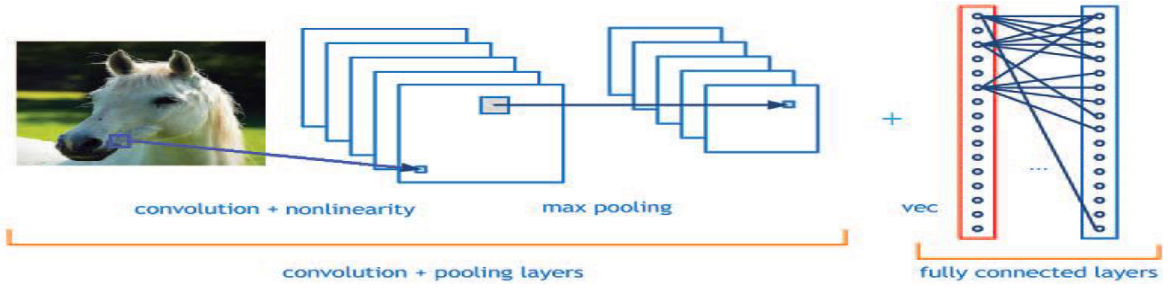


Fig.2. 5 *Une conception traditionnelle de réseaux neuronaux convolutionnels*

Sur la figure 2.6, la structure du bloc d'extraction de caractéristiques du CNN proposé est illustrée.

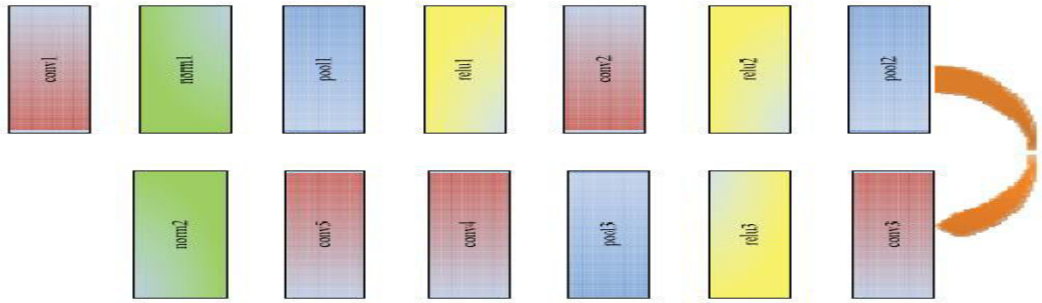


Fig.2. 6 *La structure du bloc d'extraction de caractéristiques du CNN proposé [59].*

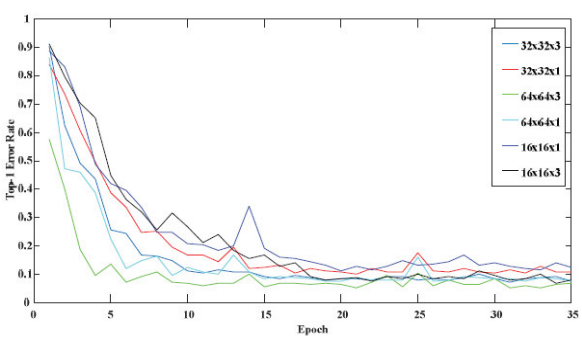
2.6.2.2 Résultats expérimentaux : le CNN a été conçu avec la version Beta23 de l'outil logiciel Mat ConvNet. Après l'étape de prétraitement, la taille de chaque image a été modifiée en 16x16x1, 16x16x3, 32x32x1, 32x32x3, 64x64x1 et 64x64x3. 66% des images ont été assignées comme ensemble de formation, 34% comme ensemble de test. On a mis en œuvre différents tests en modifiant la taille de l'image, le taux d'apprentissage, la taille des lots, etc. CNN a été formé pour 35 époques. Les performances du CNN proposé ont été évaluées en fonction des erreurs top-1 et top-5. Le taux d'erreur Top-1 vérifie si la classe supérieure est la même que l'étiquette cible et si le label cible est l'un des cinq premières prédictions. Une brève structure de l'algorithme proposé est décrite dans le tableau 1. Les résultats sont meilleurs que ceux de la littérature qui utilisent des techniques d'apprentissage superficielles.

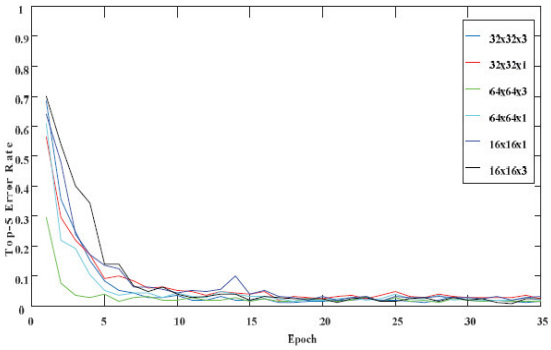
Tab.2. 5 Paramètres de l'algorithme proposé [59].

Input Image size	Number of Epoch reached at the highest rate for Top-1 Error	Number of Epoch reached at the highest rate for Top-5 Error	Batch Size	Learning Rate	Top-1 Error (Accuracy Rate)	Top-5 Error (Accuracy Rate)
16x16x1	20	24	10	0.001	88.8	98.4
16x16x3	34	21	20	0.001	93.2	98.8
32x32x1	21	17	30	0.001	90	98
32x32x3	17	31	20	0.001	92.8	98.8
64x64x1	19	18	30	0.001	92.4	98.4
64x64x3	21	28	10	0.001	94.8	98.8

La performance de l'architecture CNN proposée en termes de taux d'erreur du top-1 est illustrée sur la figure 2.8 (a). Comme le montre la figure 5, le taux d'erreur du top-1 le plus

faible a été obtenu à partir d'une image de 64x64x3. Ce résultat est important quand il est destiné à trouver l'étiquette cible de n'importe quel sujet dans la base de données. Le taux d'erreur Top-5 est présenté à la figure 2.7 (b)





et le taux le plus bas a été obtenu à partir de toutes les images à trois canaux.

(a) Top-1

(b) Top-5

Fig.2. 7 *Taux erreur du CNN proposé.*

Cet article [59] présente une évaluation empirique du système de reconnaissance faciale basé sur l'architecture CNN. Les principales caractéristiques de l'algorithme proposé sont qu'il utilise la normalisation par lots pour les sorties des première et dernière couches convolutives et que le réseau atteint des taux de précision élevés. Dans une étape de couche entièrement connectée, Softmax Classifier a été utilisé pour classer les faces. La performance de l'algorithme proposé a été testée sur Georgia Tech Face Database. Les résultats ont montré des taux de reconnaissance satisfaisants selon les études de la littérature.

2.6.3 Retracer les images vers leur réseau social d'origine: une approche basée sur CNN

Lorsqu'une image est téléchargée sur un réseau social, elle subit un traitement spécifique qui inclut généralement une compression JPEG mais aussi éventuellement redimensionnement et filtrage pour en adapter la qualité. Même si le processus réel et leurs paramètres (tels que la qualité de compression) ne sont pas connus, certaines caractéristiques distinctives sur les images sont laissées. En conséquence, ils peuvent être détectés par un classificateur. L'hypothèse est que ces caractéristiques sont principalement liées aux paramètres de compression JPEG. On choisit alors naturellement d'utiliser des fonctionnalités

basées sur DCT, car elles sont fortement affectées par compression JPEG. De plus, ils réussissent des tâches similaires telles que la détection de la double compression.

Sur cette base, un CNN basé sur le domaine fréquentiel a été conçu en prenant comme entrée une représentation statistique des coefficients DCT, et sort directement la classe du réseau social à l'origine de l'image. En supposant que K réseau social, le réseau aura donc K classes dans la sortie du réseau [61].

Considérant qu'un CNN a besoin d'une entrée de taille fixe, les auteurs ont décidé d'utiliser la quantité définie d'histogrammes de l'image coefficients DCT.

DCT est principalement affecté par le contenu et la taille de l'image considérée et afin d'être indépendant par rapport à la résolution de l'image, chaque image est subdivisée en patches non chevauchants au lieu de traiter l'ensemble image comme une seule entrée. Chaque patch est ensuite envoyé au réseau et les sorties sont finalement affinées pour obtenir la classe finale entrée [62]. Le modèle CNN proposé est basé sur des idées similaires tirées de la littérature de classification d'images [63], et son architecture est illustrée sur la figure 2.8 Chaque patch d'entrée N × N est prétraité pour calculer, comme décrit avant, le vecteur de caractéristiques qui est ensuite envoyé au réseau pour obtenir l'une des K classes de réseaux sociaux. On emploie deux blocs constitués d'un bloc convolutif unidimensionnel suivi de couches de max-pooling pour réduire la dimensionnalité et les exigences de calcul de l'approche.

Ensuite, trois couches entièrement connectées sont utilisées pour calculer la sortie finale du réseau. Chaque bloc convolutif est défini comme:

$$f(x) = g(W * x + b) \qquad (Eque\ 2.1)$$

où * est l'opérateur convolutionnel, W sont les poids 1-D de la couche, b est le biais et g est une fonction d'activation non-linéaire.

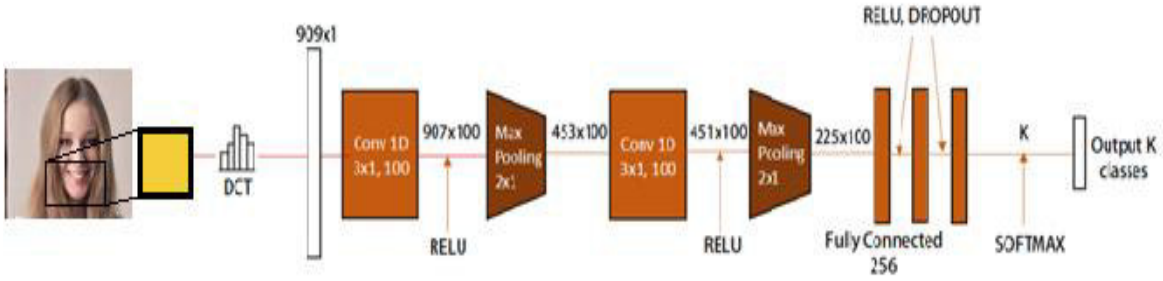


Fig.2. 8 *Architecture de la classification d'images.*

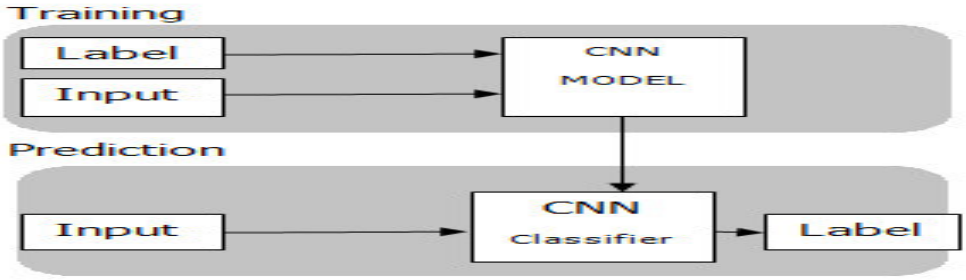
Cet article [61] propose une nouvelle méthodologie basée sur les réseaux de neurones convolutionnels (CNN) pour remonter au réseau social de provenance d'une image donnée sans recourir à ses métadonnées. La technique présentée a été testée sur trois jeux de données publics jusqu'à sept réseaux sociaux ou applications de messagerie instantanée les plus courants. Les résultats obtenus démontrent une bonne capacité de l'approche CNN proposée à distinguer entre les différentes plates-formes sociales. Des travaux futurs seront consacrés à l'augmentation du nombre de réseaux sociaux considérés, en évaluant également différents types d'architectures CNN. Un autre sujet intéressant sera de comprendre le comportement de la méthode proposée dans le cas de téléchargement multiple,

2.6.4 Reconnaissance de visage basée sur la fonctionnalité LBP pour CNN

Le travail principal de cet article [64] étudie la méthode de l'extraction faciale basée sur le modèle binaire local (LBP) et la méthode de reconnaissance faciale basée sur les réseaux de neurones à convolution (CNN). Dans cet article [64], les auteurs proposent la carte des fonctions LBP en tant que contribution de CNN pour améliorer l'apprentissage et la compréhension de CNN, ce qui fournira des pistes pour la sélection des données d'apprentissage CNN.

Le diagramme de reconnaissance de visage basé sur CNN est comme montré dans la Fig.2.9

Fig.2. 9 *Le mode de CNN identifie le visage [64]*



Méthode de reconnaissance de visage basée sur CNN, entrée CNN est une image, CNN apprend à partir du niveau de pixel. Dans le processus d'apprentissage de CNN, l'extraction de caractéristiques commence à partir de la fonctionnalité de niveau le plus bas. Il y a un apprentissage plus difficile dans l'extraction de caractéristiques.

Du point de vue de la reconnaissance, la méthode statistique locale est appliquée pour faire face à la reconnaissance basée sur la fonction LBP, et les informations de position de la carte d'entités LBP sont négligées. La fonction LPB est l'unité la plus élémentaire de la direction de la texture de l'image, et la méthode statistique pour faire correspondre les deux faces donnera des informations incomplètes.

2.6.4.1 Reconnaissance de visage basée sur la fonctionnalité LBP pour CNN : Cet article présente une méthode de reconnaissance faciale basée sur la combinaison LBP de fonctionnalités et CNN. Le CNN est entraîné en utilisant la carte de fonctionnalité LBP de l'image de visage en tant qu'entrée du CNN. Ce qui suit décrit en détail l'implémentation de la reconnaissance faciale basée sur les fonctionnalités LBP pour CNN.

2.6.4.2 Comparaison des résultats d'expériences de deux modèles : Dans le processus de conception du classificateur, l'évaluation du classificateur est très importante. Un bon indice d'évaluation est plus avantageux pour nous d'optimiser le modèle de classification; dans le même temps, un bon indice d'évaluation des classificateurs exige qu'il reflète la capacité du classificateur à résoudre le problème, et aussi plus facile de montrer l'interaction aux utilisateurs et aux clients.

Dans le problème de classification, une instance peut être déterminée comme l'un des quatre types suivants:

Exactitude: la capacité de juger l'ensemble de l'échantillon, c'est-à-dire que le jugement positif est positif et négatif le jugement est négatif.

Précision = $(TP + TN) / (TP + FN + FP + TN)$ $(Eque\ 2.2)$

Sensibilité: La capacité d'un échantillon positif à être prédite comme un échantillon positif,

Sensibilité = $TP / (TP + FN)$ $(Eque\ 2.3)$

Spécificité: La capacité des échantillons négatifs à prédire des échantillons négatifs,

Spécificité = $TN / (TN + FP)$ $(Eque\ 2.1)$

Tab.2. 6 *Tableau comparatif des deux modèles de reconnaissance de visage [64]*

Evaluation index	*Accuracy*	*Sensitivity*	*Specificity*
CNN	*91.83%*	*89.50%*	*93. 00%*
LBP+CNN	*95.33*	*90.50%*	*97.75%*

Cet article propose un modèle de reconnaissance faciale basé sur la fonctionnalité LBP et CNN. Premièrement, l'image est transformée en carte d'entités LBP, puis la carte d'entités LBP est utilisée comme entrée de CNN pour former CNN. En reconnaissance faciale et ensuite dans le CNN pour identifier en comparant l'exactitude, la sensibilité, la spécificité des deux catégories de classification, l'effet de la méthode de classification proposée est meilleur que l'effet de l'ancienne classification; selon la courbe ROC et l'analyse AUC, la méthode de reconnaissance faciale basée sur LBP et CNN est supérieure à la reconnaissance faciale CNN. En comparant les données expérimentales, on voit que, par rapport aux connaissances originales, les connaissances traitées sont plus faciles à apprendre et à comprendre par CNN,

et la reconnaissance du visage est meilleure. Par conséquent, lors de la formation CNN, on doit choisir les connaissances après le traitement.

2.6.5 Reconnaissance visage utilisant un apprentissage profond modifié

En vue d'aborder la question du faible taux de reconnaissance sur des ensembles de données plus petits dans les systèmes de reconnaissance faciale fondés sur l'apprentissage en profondeur; les auteurs [65] proposent une approche d'apprentissage en profondeur modifiée. Dans cette approche, ils utilisent la technique d'augmentation des données pour augmenter le nombre de données d'apprentissage et améliorer la puissance de généralisation du réseau. La meilleure caractéristique de cette approche est qu'elle est très simple et facile à mettre en œuvre. Les détails de l'approche sont présentés dans les étapes suivantes:

1) Sélection d'un sous-ensemble d'échantillons d'image d'un ensemble de données initialement étiqueté et ajoutez-le à l'ensemble d'apprentissage. Création d'un ensemble de tests à partir des échantillons restants.

2) Génération des images synthétiques en appliquant du bruit à chacun des échantillons d'entraînement dans l'ensemble d'apprentissage. Le bruit appliqué peut être l'un des suivants: a. Bruit de Poisson, b. Bruit gaussien.

3) Ajout de ces échantillons d'images bruitées générées synthétiquement à l'ensemble d'entraînement pour créer l'ensemble d'entraînement augmenté avec le double du nombre d'échantillons d'entraînement.

4) L'ensemble d'entraînement augmenté est ensuite fourni au modèle CNN pour la formation.

5) L'ensemble de test est ensuite reconnu avec le réseau formé ci-dessus.

6) Le pourcentage de taux de reconnaissance est calculé sur la base du nombre de correspondances correctes par rapport au nombre total d'images de test.

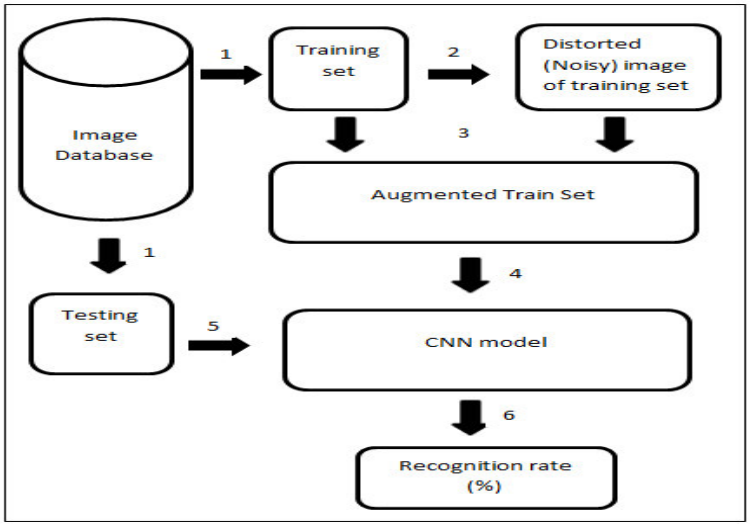


Fig.2. 10 Schéma fonctionnel de l'approche proposée [65].

Les réseaux de neurones convolutionnels profonds ont montré des performances significatives dans le domaine de la vision par ordinateur, y compris le problème difficile de la reconnaissance faciale. *Bien que la formation des modèles CNN montre des performances exceptionnelles avec un grand ensemble de données, ils ne sont pas adaptés à l'apprentissage à partir d'ensembles de données avec peu d'échantillons*.

Pour contrer ce problème d'apprentissage de la représentation faciale à partir d'un ensemble de données plus petit, une nouvelle approche est proposée où l'ensemble de données d'apprentissage est augmenté d'échantillons générés synthétiquement en ajoutant du bruit Gaussien ou de Poisson à l'ensemble d'apprentissage. Cette technique peut être appliquée à d'autres problèmes d'analyse de visage à l'avenir.

La limite de cette approche est qu'elle utilise un apprentissage supervisé avec des données humaines annotées. Les modèles d'apprentissage en profondeur sont souvent importants, ils nécessitent donc une grande quantité de mémoire et une puissance de calcul plus importante, ce qui peut être réalisé avec l'utilisation de GPU. On doit ajuster de nombreux paramètres et ne pas comprendre ou interpréter le modèle à mesure qu'il devient plus gros et plus compliqué.

Conclusion

Dans ce chapitre, nous avons minutieusement examiné plusieurs définition sur le Machine Learning et l'apprentissage en profondeur pour la reconnaissance de visage. Ensuite, nous avons présenté différentes approches de l'état de l'art sur l'application du Deep Learning pour la reconnaissance et la classification des images. Dans la dernière partie, nous avons mis en revue les travaux récents des chercheurs dans le domaine de la reconnaissance de visage basé sur l'apprentissage en profondeur et particulièrement le réseau CNN. Cet état de l'art nous a été bénéfique et nous permet de nous familiariser avec l'outil CNN et pousser nos investigations dans ce domaine pour des applications de reconnaissance de visage en milieux incontrôlés. Dans le chapitre 3, nous ferons l'étude du modèle choisi qui sera conçu et expérimenté dans le chapitre 4.

Chapitre 2 Etat de l'Art sur le Deep Learning pour la Reconnaissance

Chapitre 3

Deep Learning pour la Reconnaissance de Visage

Chapitre 3 Deep Learning pour la Reconnaissance de Visage

Introduction

Après avoir présenté les différentes méthodes de la reconnaissance faciale dans les chapitres précédents et en se basant sur les avantages et les limites de chacune d'elles. Nous nous orientons vers les méthodes d'apprentissage profond le « Deep Learning » qui préservent implicitement les informations locales sur le visage et globale et qui se basent sur un apprentissage automatique basé sur plusieurs couches de neurones. Nous choisissons la méthode CNN pour faire la reconnaissance faciale pour sa simplicité et surtout pour son aptitude à l'extension.

Dans ce chapitre, nous présentons l'essentiel et aussi une étude détaillée sur l'outil CNN et ses différentes composantes. Ce chapitre vise à fournir une présentation au concept de réseaux de neurones convolutionnels. Pour ce faire, il est nécessaire de comprendre le concept du réseau neuronal artificiel, de sorte que la première partie du chapitre lui est consacrée. Après cela, Deep Learning et CNN sont expliqués. Rappelons que, les réseaux de neurones artificiels sont des ensembles de nœuds de calcul interconnectés, généralement avec des formes carrées ou cubiques inspirés de leurs homologues biologiques. Ils constituent une approche computationnelle pour les cas dans lesquels la solution du problème, ou la recherche d'une représentation adéquate, est difficile pour les programmes informatiques traditionnels. La façon dont ils traitent l'information pourrait être comprise comme la réception d'intrants externes pouvant provoquer, ou non, une réponse dans certains des nœuds des neurones du système. L'ensemble des réponses détermine la sortie finale du réseau. Ils ont prouvé leur capacité dans de nombreux problèmes, tels que celui de la vision par ordinateur, qui sont difficiles à résoudre en extrayant les fonctionnalités de manière traditionnelle. Cette partie vise à présenter brièvement les principaux concepts techniques de la méthode, afin de faciliter la compréhension du Deep Learning.

Il existe plusieurs architectures de calcul qui tentent de modéliser le néocortex. Ces modèles ont été inspirés par certaines sources, qui tentent de cartographier diverses phases de calcul dans la compréhension de l'image à des zones dans le cortex. Au fil du temps, ces modèles ont été affinés; Cependant, le concept central du traitement visuel sur une structure hiérarchique est resté. Des organisations similaires sont utilisées par les CNN ainsi que d'autres modèles à

couches profondes (comme le Neocognitron et HMAX), mais des modèles corticaux plus «explicites» cherchent à mieux cartographier leur architecture. modèles biologiquement inspirés. En particulier, ils tentent de résoudre les problèmes d'apprentissage et d'invariance à travers divers mécanismes tels que l'analyse temporelle, dans laquelle le temps est considéré comme un élément inséparable du processus d'apprentissage. Notre motivation est justifiée par le fait que les réseaux de neurones convolutionnels (CNN) sont principalement axés sur le fait qu'ils sont bien établis dans le domaine de l'apprentissage en profondeur et qu'ils sont très prometteurs pour les travaux futurs.

3.1 Réseaux de neurones convolutionnels (CNN)

Parmi les réseaux de neurones profonds, ceux qui sont le plus largement utilisés dans les problèmes de vision par ordinateur sont les réseaux de neurones convolutionnels, basés sur l'architecture perceptron multicouche. Les NN (Neural network) suivent les mêmes principes que le cortex visuel des animaux. Cela consiste en des neurones qui ne traitent que de petites parties de l'image d'entrée - ou du champ visuel - et qui sont chargés de reconnaître les motifs pertinents. Ces neurones sont empilés dans des structures similaires à des couches, ce qui permet des modèles de plus en plus complexes. Le nombre de poids à apprendre pour une image est exceptionnellement élevé. Ceci est généralement impossible à traiter si on est confronté à une MLP *(multilayer perceptron : perceptron multicouche*) normale. Pire encore, cette connectivité complète entre les couches successives est, dans certains cas, inutile, car la position spatiale n'est pas prise en compte. Les CNN, quant à eux, se basent uniquement sur la corrélation locale, chaque neurone ne considère qu'une petite partie de son entrée dans son ensemble, et néglige le reste, économisant ainsi beaucoup d'arêtes et prenant en compte la relation entre pixels proches. En plus de cela, les éléments reconnaissables sont les mêmes indépendamment de leur position dans l'image, il est logique de chercher des bords et des coins autour de toute l'image. Par conséquent, chaque neurone recherchera les mêmes caractéristiques que le reste des neurones dans sa couche, mais dans des endroits différents. En tant que tel, le concept de poids partagés qu'ils utilisent consiste en ce qu'il n'y a qu'un seul ensemble de poids pour tous les neurones dans une couche.

Dans les réseaux CNN, de petites parties de l'image (appelées un champ réceptif local) sont traitées comme des entrées de la couche la plus basse de la structure hiérarchique. L'information se propage généralement à travers les différentes couches du réseau où, à chaque couche, un filtrage numérique est appliqué afin d'obtenir des caractéristiques saillantes

des données observées. Le procédé fournit un niveau d'invariance au déplacement, à l'échelle et à la rotation lorsque le champ réceptif local permet au neurone ou à l'unité de traitement d'accéder à des éléments élémentaires tels que des arêtes ou des coins orientés. Essentiellement, l'image d'entrée est convolutionnée avec un ensemble de N petits filtres dont les coefficients sont soit formés, soit prédéterminés en utilisant certains critères. Ainsi, la première couche (ou la couche la plus basse) du réseau est constituée de «cartes de caractéristiques» qui sont le résultat des processus de convolution, avec un biais additif et éventuellement une compression ou une normalisation des caractéristiques. Cette étape initiale est suivie d'un sous-échantillonnage (généralement une opération de moyenage de 2 3 2) qui réduit encore la dimensionnalité et offre une certaine robustesse aux déplacements spatiaux (voir la figure **3.1**). La carte des caractéristiques sous-échantillonnées reçoit ensuite un biais de pondération et de formation et se propage finalement à travers une fonction d'activation. Certaines variantes existent avec seulement une carte par couche ou des sommations de plusieurs cartes.

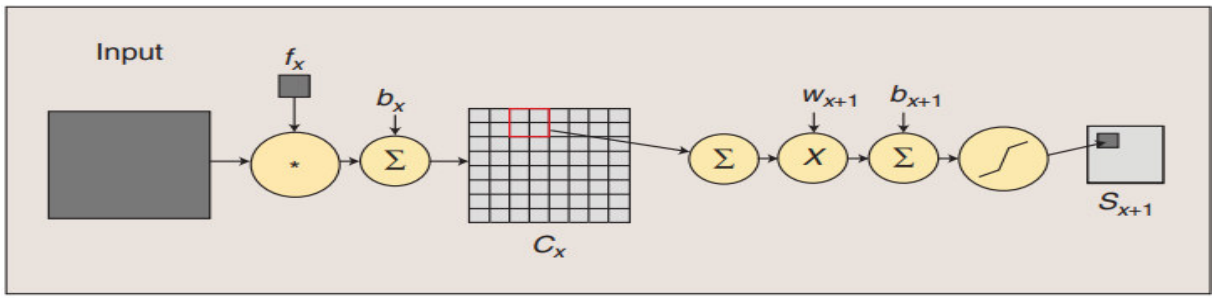


Fig.3. 1 *Processus de convolution et de sous-échantillonnage*

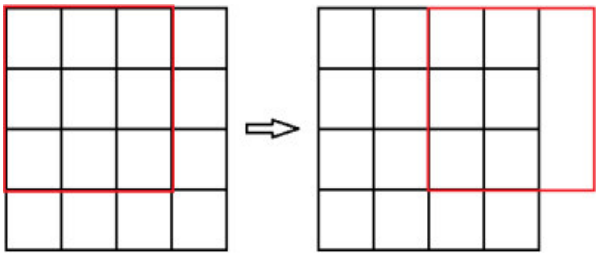
Le processus de convolution consiste à convoler une entrée (image pour la première étape ou carte de caractéristiques pour les étapes ultérieures) avec un filtre formable fx puis ajouter une polarisation bx pour former la couche de convolution Cx. Le sous-échantillonnage consiste à sommer un voisinage (quatre pixels), pondéré par scalx wx + 1, ajoutant un biais d'apprentissage bx + 1, et passant par une fonction sigmoïde pour produire une carte de caractéristiques plus ou moins 2x Sx + 1[66].

La partie de l'espace d'entrée à laquelle un neurone est connecté est appelée champ réceptif et peut se chevaucher avec celle d'autres neurones. Chaque champ réceptif est un espace 3D avec la largeur et la hauteur de celui-ci, et le nombre de canaux d'entrée. Par conséquent, pour chaque couche, il n'y a qu'un poids par valeur dans le champ réceptif. En tant que tel, si le

champ réceptif est 5 × 5 × 3, il n'y aura que 75 poids dans cette couche, utilisés par tous ses neurones. Cela permet de former de grands réseaux, le nombre de poids restant relativement faible. Cela étant dit, toutes les couches de CNN ne suivent pas ces règles. Au lieu de cela, ce type de couche s'appelle convolutional layer, mais il y en a d'autres qui seront expliqués plus loin dans cette section. Cependant, l'utilisation des couches convolutionnelles permet de réduire considérablement le nombre de poids dans le réseau, et sont les plus emblématiques de CNN. L'ensemble des poids de chaque couche est appelé le noyau, ou filtre, de cette couche. La raison en est que, lorsqu'elles sont partagées, une étape en avant de la couche peut être interprétée comme la convolution des poids et l'entrée. La fonction d'activation la plus couramment utilisée est l'unité **L**inéaire Redresseur (**Relu**), qui applique la fonction suivante à la sortie du neurone: f (x) = max (0; x). Des variations sur cette fonction sont également utilisées, telles que Noise **Relu**. Les couches de neurones sont empilées, et leurs sorties forment des volumes 3D. L'entrée de la première couche, c'est-à-dire l'image elle-même, peut également être considérée comme un volume 3D : Width× Height × Channels. Chaque pile de couches utilise une configuration de noyau différente, et toutes les couches qui la composent sont connectées à toutes les couches de la pile précédente. Les principaux paramètres à prendre en compte dans les couches CNN sont:

- ***Taille du noyau:*** la largeur et la hauteur du champ réceptif pour les neurones de cette pile. Les deux côtés sont généralement de taille égale - des grains carrés.
- ***Stride:*** chaque neurone traite une région de l'espace d'entrée. Comme ces régions peuvent se chevaucher, la pooling indique la distance entre leurs centres. En tant que tel, une pooling de 1 signifie que chaque neurone traite la même région que son voisin à l'exception d'une colonne. Plus la pooling est grande, plus la largeur et la hauteur de sortie de cette couche sont petites.

- ***Padding :*** Dans certains cas, les neurones situés dans les limites de la couche ne peuvent pas traiter tout un champ réceptif. Cela peut arriver en raison de la foulée.



Par exemple, dans la figure 3.3, nous avons un exemple d'une image 4 × 4 traitée par un noyau 3 × 3 avec un pooling de 2. Comme la différence entre les champs réceptifs est de 2 pixels, la dernière colonne du champ réceptif du second neurone tombe en dehors de l'image et n'est pas traitée. Afin de résoudre ce problème, une possibilité est d'ajouter une "bordure" autour de l'image de 0. De cette façon, nous garantissons que tous les neurones traitent un champ réceptif. Il s'agit du padding ou le rembourrage, s'il est utilisé, est généralement 1 ou 2.

- ***Type de couche:*** Ceci sera expliqué plus en détail par la suite, mais ici nous disons simplement que toutes les couches d'une pile appartiennent au même type. Les plus

Fig.3. 2 *Exemple de cas justifiant le Padding*

communément utilisés sont les couches de regroupement, de convolution et de connexion complète. Ce sont les paramètres les plus utilisés, bien que tous les types de réseaux ne les utilisent pas tous. Ces paramètres ont une conséquence intéressante: même si dans les NN normales, le nombre de neurones de chaque couche est spécifié, il ne l'est pas dans les CNN. Au lieu de cela, il est déduit des paramètres de la couche, et il consiste en le nombre de neurones nécessaires pour traiter toute l'image. Par conséquent, la taille du côté de la couche est obtenue en utilisant l'équation 3.1 $H = \frac{W-F+2*P}{S+1}$ Equ 3.1

Où W est le côté de l'espace d'entrée, F le côté du noyau utilisé, P correspond au remplissage et S à la chaîne. Par conséquent, la sortie de toutes les couches d'une pile sera la même. Ceci, avec le nombre de couches empilées, indique la taille de la sortie de cette pile. Cela étant dit, la sortie d'une couche ne sera jamais plus grande en largeur et en hauteur que son entrée, et le

plus souvent elle sera plus petite. Le paramètre de pooling, en particulier, a le potentiel d'effectuer des réductions importantes, car une pooling de 2 réduira de moitié la taille du côté de l'image. Par conséquent, lorsque l'image est propagée à travers le réseau, elle devient plus petite (figure. 3.3), Cela signifie que chaque neurone des dernières couches traitera des taches plus grandes de l'image originale.

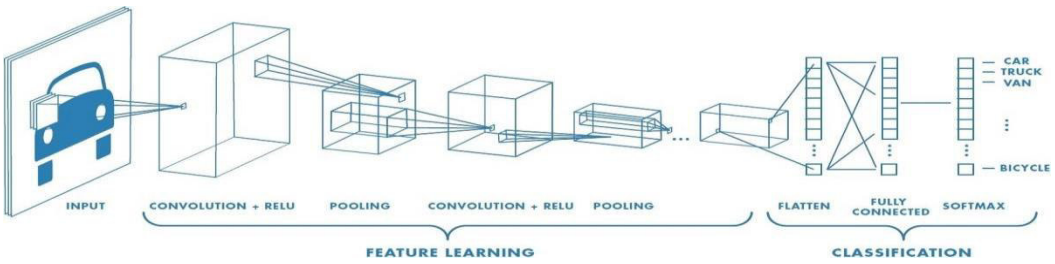


Fig.3. 3 *Exemple d'architecture CNN [67]*

Ici nous pouvons voir chaque pile de couches et comment chaque neurone ne traite qu'un patch de son entrée. La taille de l'image continue à diminuer jusqu'à atteindre les couches entièrement connectées.

3.2 Spécification des couches du réseau neuronal convolutionnel

La première étape de la création et de la formation d'un nouveau réseau de neurones convolutif (ConvNet) consiste à définir l'architecture du réseau. Cette rubrique explique les détails des couches ConvNet et l'ordre dans lequel elles apparaissent dans un réseau ConvNet. L'architecture d'un réseau ConvNet peut varier en fonction des types et des nombres de couches incluses. Le modèle de la figure 3.4 illustre une structure de couche avec des paramètres d'apprentissage.

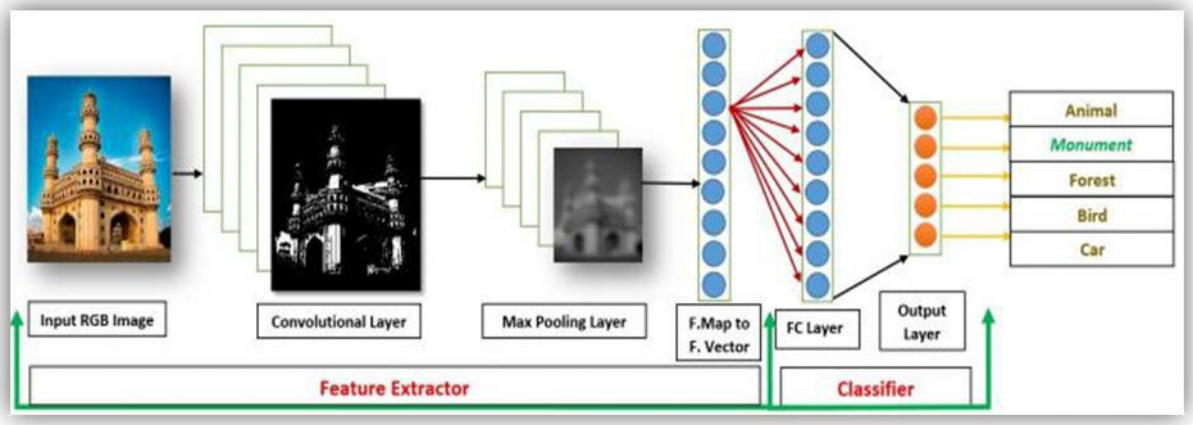


Fig.3. 4 *Architecture du réseau de neurones convolutionnel [68].*

Définition des couches d'apprentissage en profondeur : On doit définir une couche d'apprentissage en profondeur personnalisée et spécifier des paramètres d'apprentissage facultatifs, des fonctions de transfert et une fonction de retour.

Les types et le nombre de couches incluses dépendent de l'application ou des données particulières. Par exemple, si on a des réponses catégoriques, on doit avoir une couche softmax et une couche de classification, alors que si la réponse est continue, on doit avoir une couche de régression à la fin du réseau. Un réseau plus petit avec seulement une ou deux couches de convolution peut être suffisant pour apprendre sur un petit nombre de données d'image en niveaux de gris. D'un autre côté, pour des données plus complexes avec des millions d'images colorées, vous aurez peut-être besoin d'un réseau plus complexe avec plusieurs couches convolutives et entièrement connectées.

3.2.1 Couche d'image d'entrée (Image Input Layer)

La couche d'entrée d'image définit la taille des images d'entrée d'un réseau de neurones convolutives et contient les valeurs de pixels brutes des images. On peut ajouter une couche d'entrée à l'aide de la fonction ***imageInputLayer*** et spécifier la taille de l'image à l'aide de l'argument ***inputSize***. La taille d'une image correspond à la hauteur, la largeur et le nombre de canaux de couleur de cette image.

Remarque : *Pour une image en niveaux de gris, le nombre de canaux est 1, et pour une image couleur, il est 3.*

3.2.2 Couche convolutionnelle (Convolutional Layer)

La couche la plus emblématique a déjà été introduite. C'est inspiré des MLP traditionnels, mais avec des différences majeures. Les principales sont que chaque couche a un seul ensemble de poids pour tous les poids partagés des neurones, et que chaque neurone ne traite qu'une petite partie de l'espace d'entrée. Il utilise tous les paramètres introduit dans la section précédente.

3.2.3 Couche de regroupement (Pooling Layer)

Un autre concept important des CNNs est le pooling, ce qui est une forme de sous-échantillonnage de l'image. L'image d'entrée est découpée en une série de rectangles de n pixels de côté ne se chevauchant pas (pooling). Chaque rectangle peut être vu comme une tuile. Le signal en sortie de tuile est défini en fonction des valeurs prises par les différents

pixels de la tuile. Le pooling réduit la taille spatiale d'une image intermédiaire, réduisant ainsi la quantité de paramètres et de calcul dans le réseau. Il est donc fréquent d'insérer périodiquement une couche de pooling entre deux couches convolutives successives d'une architecture de réseau de neurones convolutives pour réduire le sur apprentissage. La couche de pooling fonctionne indépendamment sur chaque tranche de profondeur de l'entrée et la redimensionne uniquement au niveau de la surface. La forme la plus courante est une couche de mise en commun avec des tuiles de taille 2 × 2 (largeur/hauteur) et comme valeur de sortie la valeur maximale en entrée Son utilité consiste à réduire la quantité de poids à apprendre, ce qui réduit le temps de calcul ainsi que la probabilité de sur apprentissage.

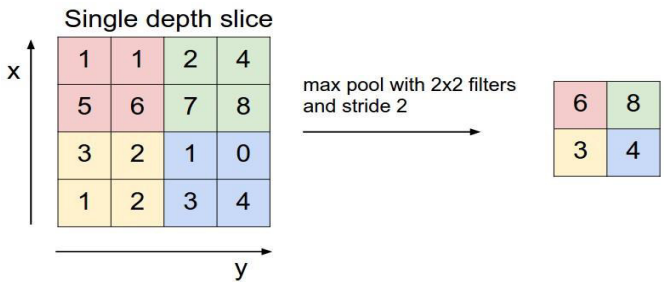


Fig.3. 5 *Max pooling avec un filtre 2 × 2 et un pas de 2 [69].*

.

3.2.4 Couches de correction (Relu)

Souvent, il est possible d'améliorer l'efficacité du traitement en intercalant entre les couches de traitement une couche qui va opérer une fonction mathématique (fonction d'activation) sur les signaux de sortie. On a notamment : La correction **Relu** (abréviation de Unité Linéaire Rectifiée). Cette fonction, appelée aussi « fonction d'activation non saturante », augmente les propriétés non linéaires de la fonction de décision et de l'ensemble du réseau sans affecter les champs récepteurs de la couche de convolution [70].

La correction par tangente hyperbolique : $f(x) = \tanh(x)$ Equ 3.2

La correction par la tangente hyperbolique saturante : $f(x) = |\tanh(x)|$ Equ 3.3

La correction par la fonction sigmoïde : $f(x) = (1 + e^{-x})^{-1}$ Equ 3.3

3.2.5 Couche entièrement connectée (Fully Connected Layer (FC))

Après plusieurs couches de convolution et de max-pooling, le raisonnement de haut niveau dans le réseau neuronal se fait via des couches entièrement connectées. Les neurones dans une couche entièrement connectée ont des connexions vers toutes les sorties de la couche précédente (comme on le voit régulièrement dans les réseaux réguliers de neurones). Leurs fonctions d'activations peuvent donc être calculées avec une multiplication matricielle suivie d'un décalage de polarisation. Ces couches sont essentiellement des couches neurales reliées à tous les neurones de la couche précédente. Dans ce cas, elles n'utilisent aucun des paramètres introduits, en utilisant à la place le nombre de neurones. La sortie qu'ils produisent peut être comprise comme un vecteur de caractéristiques compact représentant l'image d'entrée. Elles sont également utilisées comme couches de sortie, avec un neurone par sortie, comme d'habitude.

3.2.6 Couche de sortie de classification

On doit définir une couche de sortie de classification personnalisée et spécifier une fonction de perte.

3.2.7 Architecture de couche (layer)

Une couche a deux composants principaux : le passage avant et le passage arrière. Pendant le passage avant d'un réseau, la couche prend la sortie x de la couche précédente, applique une fonction, puis émet (propage en avant) le résultat z vers la couche suivante. A la fin du passage avant, le réseau calcule la perte L entre les prédictions Y et la vraie cible T. Lors du passage en arrière d'un réseau, chaque couche prend les dérivées de la perte par rapport à z, calcule les dérivées de la perte L par rapport à x, puis émet des résultats (se propage vers l'arrière) à la couche précédente. Si la couche a des paramètres apprenants, alors la couche calcule également les dérivées des pondérations de couche (paramètres apprenables) W. La couche utilise les dérivées des poids pour mettre à jour les paramètres apprenables. La figure **3. 1** suivante décrit le flux de données à travers un réseau neuronal profond et met en évidence le flux de données à travers la couche.

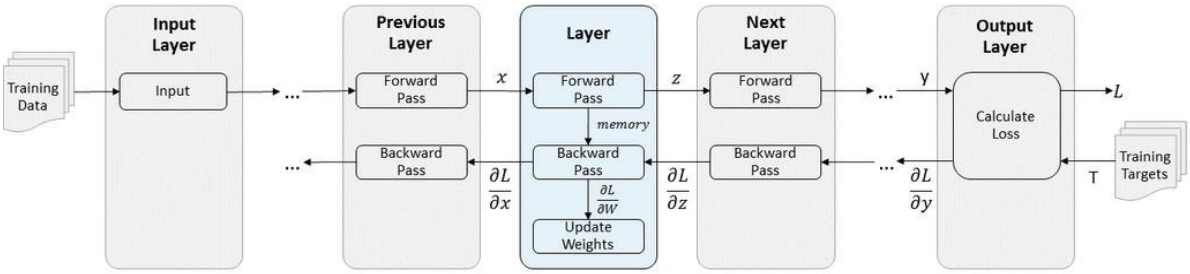


Fig.3. 6 *Description du flux de données à travers un réseau neuronal profond et met en évidence le flux de données à travers la couche [71].*

3.2.8 Propriétés de couche (layer)

Les propriétés de la classe sont définies en définissant la classe par défaut. Les couches définies par l'utilisateur contiennent trois propriétés:

Nom : le nom de la classe est défini comme un vecteur de caractères. Utiliser la propriété *Name* pour définir les couches et les indexer dans une grille. Si on ne spécifie pas le nom de la classe, le programme en affecte automatiquement un au moment de la formation.

Description : décrit une seule ligne de la couche et est définie comme un vecteur de caractères. Cette description apparaît lorsque l'on affiche la couche dans un tableau de couches. La *Valeur* par défaut est le nom de la classe de couche.

Type de couche : spécifiée en tant que chaine de caractères. Une valeur *Type* apparaît lorsque la couche est affichée dans un tableau de couche. La valeur par défaut est le nom de la classe de couche si la couche n'a pas d'autres propriétés, on peut supprimer la section des propriétés.

3.3 Choix des hyper paramètres

Les réseaux de neurones convolutifs utilisent plus des hyper paramètres qu'un perceptron multicouche standard. Même si les règles habituelles pour les taux d'apprentissage et des constantes de régularisation s'appliquent toujours, il faut prendre en considération les notions de nombre de filtres, leur forme et la forme du max pooling.

3.3.1 Nombre de filtres

Comme la taille des images intermédiaires diminue avec la profondeur du traitement, les couches proches de l'entrée ont tendance à avoir moins de filtres tandis que les couches

plus proches de la sortie peuvent en avoir davantage. Pour égaliser le calcul à chaque couche, le produit du nombre de caractéristiques et le nombre de pixels traités est généralement choisi pour être à peu près constant à travers les couches.

3.3.2 Forme du filtre

Les formes de filtre varient grandement dans la littérature. Ils sont généralement choisis en fonction de l'ensemble de données. Les meilleurs résultats sur les images (28 x 28) sont habituellement dans la gamme de 5 × 5 sur la première couche, tandis que les ensembles de données d'images naturelles (souvent avec des centaines de pixels dans chaque dimension) ont tendance à utiliser de plus grands filtres de première couche de 12 × 12, voire 15 × 15.

Le défi est donc de trouver le bon niveau de granularité de manière à créer des abstractions à l'échelle appropriée et adaptée à chaque cas

3.3.3 Forme du Max Pooling

Les valeurs typiques sont 2 × 2. De très grands volumes d'entrée peuvent justifier un pooling 4 × 4 dans les premières couches. Cependant, le choix de formes plus grandes va considérablement réduire la dimension du signal, et peut entraîner la perte de trop d'information.

3.4 L'outil Deep Learning ?

À travers des notes sur «l'apprentissage en profondeur» par Ian **Goodfellow, Yoshua Benjio et Aaron Corville**. L'apprentissage automatique est une branche de la statistique qui utilise des échantillons pour approximer les fonctions. Nous avons une vraie fonction ou distribution sous-jacente qui génère des données, mais nous ne savons pas ce que c'est. Nous pouvons échantillonner cette fonction, et ces échantillons forment nos données d'entraînement. Exemple de sous-titrage d'image : Fonction: f * (image) → description

Echantillons : Données ∈ (image, description)

Le but de la machine est de trouver des modèles ayant les caractéristiques suivantes :

- *Avoir assez de pouvoir de représentation pour se rapprocher de la vraie fonction ;*
- *Avoir un algorithme efficace qui utilise des données d'apprentissage pour trouver de bonnes approximations de la fonction ;*
- *L'approximation doit généraliser pour renvoyer de bons résultats pour les entrées invisibles. Applications possibles de l'apprentissage automatique:*

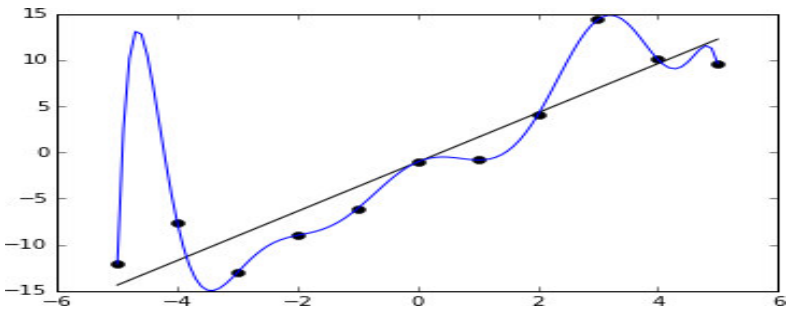
- *Convertissez les entrées en une autre forme - apprenez l'information, extrayez-la et exprimez-la. par exemple: classification d'image, sous-titrage d'image.*
- *Prédisez les valeurs manquantes ou futures d'une séquence - apprenez la «causalité» et prédisez-la.*
- *Synthétisez des résultats similaires - apprenez la «structure» et générez-la.*

.

a) Généralisation et sur-apprentissage :

Le sur-apprentissage est quand vous trouvez un bon modèle des données d'entraînement, mais ce modèle ne se généralise pas. Par exemple: un élève qui a mémorisé les réponses à des tests d'entraînement obtiendra de bons résultats à un test d'entraînement, mais pourrait avoir de mauvais résultats au test final. Il faut mettre de côté les «données de test» qui ne sont jamais entraînées. Une fois la formation terminée, nous exécutons le modèle sur les données de test finales.

On ne peut pas modifier le modèle après le test final (bien sûr, on peut rassembler plus de données). Si la formation du modèle se déroule par étapes, on doit retenir les données de test pour chaque étape. L'apprentissage en profondeur est une branche des techniques



d'apprentissage automatique. C'est un modèle puissant qui a également réussi à généraliser.

b) Réseaux Feedforward :

Fig.3. 7 *Illustration de la fonction d'approximation statistique des données (image)*

Les réseaux Feedforward représentent : $y = f^*(x)$ $\min_{\theta \in models} J\left(y, f\left(¨x;\theta\right)\right)$ Equ 3.*4*

Avec une famille de fonctions: $u = f(x;\theta)$

Equ 3.5

θ : sont les paramètres du modèle. Cela pourrait être des milliers ou des millions de paramètres $\theta_1 \ldots\ldots \theta_T$; $f(x;\theta)$ est une fonction unique de x ; u : est la sortie du modèle. On peut imaginer si on a choisi une famille de fonctions suffisamment générale, les chances sont, l'un d'eux ressemblera f *

Par exemple : les paramètres représentent une matrice et un vecteur :

$$f(\vec{x};\theta)() = \begin{bmatrix} \theta_0 & \theta_1 \\ \theta_2 & \theta_3 \end{bmatrix} \vec{x} + \begin{bmatrix} \theta_4 \\ \theta_5 \end{bmatrix} \quad \text{Equ } 3.6$$

Conception de la couche de sortie : La couche de sortie la plus courante est:

$$f(x; Mx+b) = g(Mx+b) \quad \text{Equ } 3.7$$

La partie linéaire $Mx+b$ garantit que la sortie dépend de toutes les entrées.

La partie non linéaire $g(x)$ permet d'ajuster la distribution. Par exemple pour l'entrée de photos, la distribution de sortie pourrait être:

Linéaire: par exemple, la gentillesse dans la photo.

Sigmoïde: par exemple, la probabilité est un chat.

Soft max: par ex. la probabilité est l'une des races de chats. Pour s'assurer que la distribution correspond on peut utiliser:

Par exemple pour l'entrée de photos, la distribution de sortie pourrait être:

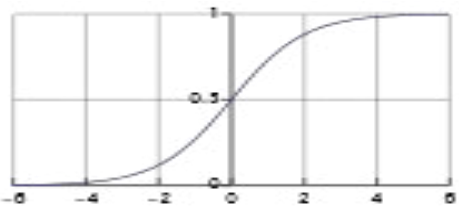


Fig.3. 8 *La distribution de sortie d'une photo*

Linéaire: $Mx+b$ Equ 3.*8*

Sigmoïde: $g(x)=\frac{1}{1+e^{-x}}$ Equ 3.*9*

Soft max est réellement sous-contraint, et souvent est mis à 1. Dans ce cas, sigmoïde est juste soft max dans 2 variables. Il y a une théorie derrière la raison pour laquelle ce sont de bons choix, mais il y a beaucoup de choix différents.

Trouver θ : Trouver θ en résolvant le problème d'optimisation suivant pour g la fonction de coût: $\min_{\theta\in models} J\left(y, f\left(\ddot{x};\theta\right)\right)$ Equ 3.*10*

L'apprentissage en profondeur est un succès car il existe une bonne famille d'algorithmes pour calculer min. Intuitivement, à chaque θ, on choisit la direction qui réduit le plus le coût. Cela nous oblige à calculer le gradient. On ne veut pas que le dégradé soit proche parce que l'on apprend trop lentement ou pas ce n'est pas stable. C'est un algorithme glouton, et pourrait donc converger mais dans un minimum local.

Choisir la fonction de coût : Cette fonction de coût pourrait être n'importe quoi :

Somme des erreurs absolues: $J=\sum|y-u|$ Equ 3.*11*

Somme des erreurs carrées: $J=\sum(y-u)^2$ Equ 3.12

Tant que le minimum se produit lorsque les distributions sont les mêmes, en théorie, cela fonctionnerait. Une bonne idée est que u représente les paramètres de la distribution de y. *Justification:* les processus naturels sont souvent flous et toute entrée peut avoir une gamme de résultats. Cette approche donne également une mesure précise de notre exactitude. Le principe du maximum de vraisemblance dit que : $\theta_{ML}=\arg\max_\theta p(y;u)$.

On veut donc minimiser : $J=-p(y;u)$ Equ 3.1*3*

c) Régularisation

Les techniques de régularisation sont des méthodes qui tentent de réduire les erreurs de généralisation. Ce n'est pas destiné à améliorer l'erreur d'entraînement. Prendre de préférence des valeurs θ plus petites: en ajoutant une fonction de θ dans j,on peut encourager de petits paramètres.

$$L^2 : J' = J + \sum|\theta|^2$$
$$L^1 : J' = J + \sum|\theta| \qquad \text{Equ 3.14}$$

Concevoir des couches cachées. (Hidden Layers).

La couche cachée la plus commune est: $f^n(x) = g(Mx+b)$. Les couches cachées ont la même structure que la couche de sortie, cependant, les $g\ (x)$qui fonctionnent bien pour la couche de sortie ne fonctionnent pas bien pour les couches cachées. Le plus simple et le plus réussi g est l'unité linéaire rectifiée (ReLU): $g\ (x) = \max(0,x)$ **Equ 3.15**

Par rapport à sigmoïde, les gradients de ReLU ne s'approchent pas de zéro quand x est très grand. D'autres fonctions non linéaires courantes incluent:

ReLU modulé: $g(x) = \max(0,x) + \alpha \min(0,x)$ Où α ,est -1, très petit ou un paramètre de modèle lui-même. L'intuition est que cette fonction a une pente pour $x < 0$. En pratique, il n'y a pas de gagnant absolu entre ceci et Relu. Max out: $g(x) = \max_{j\in} G_{(i)}xj$

d) Méthodes d'optimisation

Les méthodes utilisées sont basées sur la descente de gradient stochastique:

Il faut choisir un sous-ensemble des données d'entraînement (un ***minibatch***), et l'on calcule le gradient à partir de cela.

Avantage: ne dépend pas de la taille de l'ensemble d'entraînement, mais de la taille de la ***minibatch***.

Il y a plusieurs façons de faire une descente en dégradé (en utilisant: gradient).

Descente en dégradé - gradient d'utilisation: $\Delta = \in g$

Momentum - utiliser un gradient décroissant exponentiel: $\Delta = \in \sum e^{-t} gt$

Taux d'apprentissage adaptatif où $\in = \in_t$

AdaGrad - apprentissage lent sur une amplitude de gradient $\in_t = \frac{\in}{\delta + \sqrt{\sum g^2{}_t}}$

Simplifier le réseau : À ce point, on a suffisamment de base pour concevoir et optimiser les réseaux profonds. Cependant, ces modèles sont très généraux et de grande taille. Si le réseau a N couches chacune avec S entrées / sorties, l'espace de paramètre est $|\theta| = O(NS^2)$

e) Neurones sigmoïdes

Les algorithmes d'apprentissage semblent terribles. Mais comment peut-on concevoir de tels algorithmes pour un réseau de neurones? Supposons que nous ayons un réseau de perceptrons que nous aimerions utiliser pour apprendre à résoudre un problème. Par exemple, les entrées du réseau peuvent être les données de pixels brutes provenant d'une image numérisée et manuscrite d'un chiffre. Et nous aimerions que le réseau apprenne les poids et les biais afin que la sortie du réseau classe correctement le chiffre. Pour voir comment l'apprentissage pourrait fonctionner, supposons que nous fassions un petit changement de poids (ou de biais) dans le réseau. Ce que nous aimerions, c'est que ce léger changement de poids ne provoque qu'une petite modification correspondante de la sortie du réseau [72]

3.5 Problème de RV et Apprentissage des compétences approfondies

La tâche de reconnaissance des visages, en particulier en dehors des conditions contrôlées, est un problème extrêmement difficile. En fait, il y a eu de nombreuses approches à travers l'histoire qui n'ont pas réussi. En dehors de la variance entre les images d'un même visage, telles que l'expression, les conditions de lumière ou les poils du visage, il est difficile de déterminer ce qui rend un visage reconnaissable. En fin de compte, nous avons décidé de nous

concentrer sur les outils d'AI le Deep Learning particulièrement le CNN. Les principales raisons sont les bons résultats obtenus de l'état de l'art sur la reconnaissance d'objets et la classification d'images, et la qualité de la description.

Dans ce domaine on apprend des caractéristiques avec des variations de réseaux de neurones convolutionnels profonds (ConvNets profonds). Les opérations de convolution et de regroupement dans les ConvNets profonds sont spécialement conçues pour extraire des entités visuelles de manière hiérarchique, depuis les fonctionnalités locales de bas niveau jusqu'aux fonctions globales de haut niveau. Les ConvNets profonds adoptent des structures contenant contient plusieurs couches convolutives, avec un partage de poids local dans certaines couches convolutives. Le ConvNet extrait un vecteur de caractéristiques DeepID (Deep Identifiant) de dimension N dimensions à sa dernière couche (couche DeepID) de la cascade d'extraction de caractéristiques. La couche DeepID à apprendre est entièrement connectée aux couches convolutives. Des unités linéaires rectifiées (ReLU) sont utilisées pour les neurones dans les couches convolutives et la couche DeepID. Une illustration de la structure ConvNet utilisée pour extraire les fonctions DeepID2 (DeepID) est montrée sur la figure 3.9 avec une entrée RVB de taille 55 × 47.

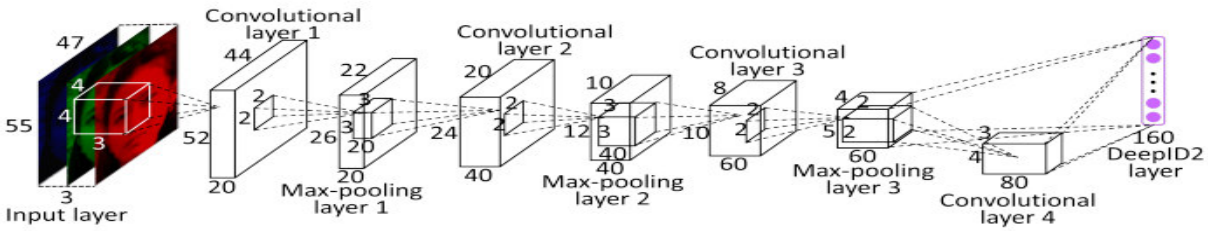


Fig.3. 9 *La structure ConvNet pour l'extraction DeepID2 [73]*

Lorsque la taille de la région d'entrée change, les tailles de carte dans les couches suivantes changent en conséquence. Le processus d'extraction de caractéristiques DeepID2 est noté :

f = Conv (x; θc), où Conv (•) est la fonction d'extraction de caractéristiques définie par le ConvNet, x est le patch de face d'entrée, f le vecteur de caractéristiques DeepID2 extrait et θc paramètres de ConvNet à apprendre.

3.5.1 Structure ConvNet pour l'extraction de fonctionnalités DeepID2 *:* Les fonctionnalités DeepID2 sont apprises avec deux signaux de supervision. Le premier est un signal d'identification de visage, qui classe chaque image de visage en une identité de n (par exemple, n = 8 192) différentes identités. L'identification est réalisée en suivant la couche DeepID2 avec une couche Softmax n-way, qui produit une distribution de probabilité sur les n classes. Le réseau est formé pour minimiser la perte d'entropie croisée, qui est appelée la perte d'identification.

où f est le vecteur de caractéristiques DeepID2, t est la classe cible et θ_{id} désigne les paramètres de la couche Softmax, est la distribution de probabilité cible, où $p_i = 0$ pour tout i sauf $pt = 1$ pour la classe cible t. est la distribution de probabilité prédite. Pour classer correctement toutes les classes simultanément, la couche DeepID2 doit former des caractéristiques discriminantes liées à l'identité (c'est-à-dire des caractéristiques avec de grandes variations interpersonnelles). Le second est le signal de vérification du visage, qui encourage les entités DeepID2 extraites des visages de la même identité à être similaires. Le signal de vérification régularise directement les caractéristiques de DeepID2 et peut réduire efficacement les variations intra-personnelles. Les contraintes couramment utilisées incluent la norme L1 / L2 et la similarité de cosinus. On [73] adopte la fonction de perte suivante basée sur la norme L2, pour la réduction de la dimensionnalité. Plusieurs architectures de réseau de neurones convolutionnelle profonde ont été développées pour la classification et la détection dans le challenge de reconnaissance visuelle à grande échelle. Dans ce qui suit nous présentons quelques une adaptée à la vision.

3.5.2 Deep Learning Image Classification AlexNet *:* *AlexNet* est un modèle pré-entraîné. Ce modèle est formé sur un sous-ensemble de la base de données Imagnet [74], qui est utilisée dans le challenge de reconnaissance visuelle à grande échelle Imagent (ILSVRC) [75]. Le modèle est formé sur plus d'un million d'images et peut classer les images en 1000 catégories d'objets. Par exemple, clavier, souris, crayon et de nombreux animaux. En conséquence, le modèle a appris de riches représentations de caractéristiques pour un large éventail d'images. *AlexNet* a été formé sur plus d'un million d'images et peut classer les images en 1000 catégories d'objets (comme le clavier, la tasse à café, le crayon et de nombreux animaux). Le réseau a appris de riches représentations de caractéristiques pour un large éventail d'images.

Le réseau prend une image comme entrée et sort une étiquette pour l'objet dans l'image avec les probabilités pour chacune des catégories d'objets. L'apprentissage par transfert est couramment utilisé dans les applications d'apprentissage en profondeur. On peut utiliser un réseau pré-entraîné et l'utiliser comme point de départ pour apprendre une nouvelle tâche. La mise au point d'un réseau avec apprentissage par transfert est généralement beaucoup plus rapide et plus facile que l'apprentissage d'un réseau avec des poids initialisés de manière aléatoire. On peut transférer rapidement des fonctions apprises à une nouvelle tâche en utilisant un nombre réduit d'images d'apprentissage.

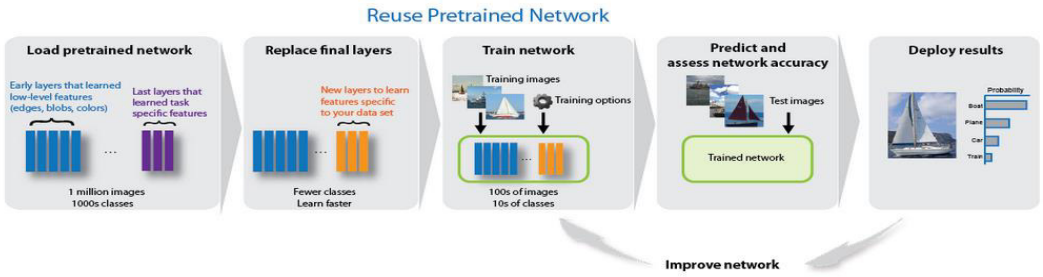


Fig.3. 10 *Exemple montrant comment affiner un CNN AlexNet pré entraîné pour une classification sur une nouvelle collection d'images [76].*

A. Classer une image à l'aide d'AlexNet

Le modèle pré-entraîné nécessite que la taille de l'image soit la même que la taille d'entrée du réseau. Déterminez la taille d'entrée du réseau à l'aide de la propriété Input Size de la première couche du réseau. Dans la mise en œuvre originale d'*AlexNet*, pour diviser le réseau entre deux GPU à mémoire limitée pour la formation, certaines couches convolutionnelles utilisent des groupes de filtres. Dans ces couches, les filtres sont divisés en deux groupes. La couche divise l'entrée en deux sections le long de la dimension du canal, puis applique chaque groupe de filtres à une section différente. La couche concatène ensuite les deux sections résultantes ensemble pour produire la sortie. Par exemple, dans la deuxième couche convolutionnelle d'*Alex Net*, la couche divise les poids en deux groupes de 128 filtres. Chaque filtre a 48 canaux. L'entrée de la couche a 96 canaux et est divisée en deux sections avec 48 canaux. La couche applique chaque groupe de filtres à une section différente et produit deux sorties avec 128 canaux. La couche concatène alors ces deux sorties pour donner une sortie finale avec 256 canaux.

Fig.3. 11 *Quelques exemples de classification d'images [77].*

```
ans =
  25x1 Layer array with layers:

     1   'data'     Image Input                 227x227x3 images with 'zerocenter' normalization
     2   'conv1'    Convolution                 96 11x11x3 convolutions with stride [4  4] and padding [0  0  0  0]
     3   'relu1'    ReLU                        ReLU
     4   'norm1'    Cross Channel Normalization cross channel normalization with 5 channels per element
     5   'pool1'    Max Pooling                 3x3 max pooling with stride [2  2] and padding [0  0  0  0]
     6   'conv2'    Convolution                 256 5x5x48 convolutions with stride [1  1] and padding [2  2  2  2]
     7   'relu2'    ReLU                        ReLU
     8   'norm2'    Cross Channel Normalization cross channel normalization with 5 channels per element
     9   'pool2'    Max Pooling                 3x3 max pooling with stride [2  2] and padding [0  0  0  0]
    10   'conv3'    Convolution                 384 3x3x256 convolutions with stride [1  1] and padding [1  1  1  1]
    11   'relu3'    ReLU                        ReLU
    12   'conv4'    Convolution                 384 3x3x192 convolutions with stride [1  1] and padding [1  1  1  1]
    13   'relu4'    ReLU                        ReLU
    14   'conv5'    Convolution                 256 3x3x192 convolutions with stride [1  1] and padding [1  1  1  1]
    15   'relu5'    ReLU                        ReLU
    16   'pool5'    Max Pooling                 3x3 max pooling with stride [2  2] and padding [0  0  0  0]
    17   'fc6'      Fully Connected             4096 fully connected layer
    18   'relu6'    ReLU                        ReLU
    19   'drop6'    Dropout                     50% dropout
    20   'fc7'      Fully Connected             4096 fully connected layer
    21   'relu7'    ReLU                        ReLU
    22   'drop7'    Dropout                     50% dropout
    23   'fc8'      Fully Connected             1000 fully connected layer
    24   'prob'     Softmax                     softmax
    25   'output'   Classification Output       crossentropyex with 'tench' and 999 other classes
```

Le réseau comporte 5 couches convolutives et 3 couches entièrement connectées.

B. Classer une image à l'aide de GoogleNet

L'une des nouvelles fonctionnalités qui a attiré notre attention est que les activations de la couche de calcul ont été étendues à GoogleNet et à Inception-v3.

Inception-v3 : une architecture de réseau de neurones convolutionnelle profonde pour la classification et la détection dans le challenge de reconnaissance visuelle à grande échelle ImageNet 2014 (ILSVRC14)

GoogleNet : un réseau de 22 couches dont la qualité est évalué dans le contexte de la classification et de la détection.

Conclusion

Dans ce chapitre nous avons mis en revue les fondements essentiels des réseaux de neurones convolutifs. Nous avons défini l'architecture détaillée de ce type de réseau ainsi que les différentes couches le constituant. Une partie a été consacrée aux propriétés et paramètres du CNN. Ensuite, nous avons présenté le principe et l'outil mathématique utile pour la compréhension et la conception de ce type d'architecture d'apprentissage profond. Cette étude détaillée du modèle CNN nous permet de mieux comprendre son processus et faire une conception pour la reconnaissance du visage qui fera objet du chapitre 4.

Chapitre 3 Deep Learning pour la Reconnaissance de Visage

Chapitre 4

Conception et Résultats du SRV basé sur le CNN

Introduction

Chapitre 4 Conception et Résultats du SRV basé sur le CNN

Dans ce chapitre, nous présentons la conception d'un système de reconnaissance automatique de visage avec l'outil CNN (Convolutional Neural Network). Grâce à notre lecture des propositions de l'état de l'art et en examinant les propositions de certains chercheurs qui stipulent une architecture CNN pour la reconnaissance faciale à travers plusieurs ensembles de données de visage standard bien connus [16]

Ce projet s'intéresse à une technique efficace pour le système de reconnaissance faciale basée sur Deep Learning utilisant CNN (Convolutional Neural Network).

Le principe général des **R**éseaux de **N**eurones **A**rtificiels (**RNA**) est à l'origine inspiré de certaines fonctions de base des neurones naturels du cerveau. Un réseau de neurones artificiel est généralement organisé en **plusieurs couches** : une couche d'entrée, une couche de sortie, des couches intermédiaires appelées couches cachées.

La présence de couches cachées permet de **discriminer** des classes d'objets non linéairement séparables. En général, un réseau de neurones est fondamentalement un classifieur. Il réalise un travail de classification pendant la phase d'apprentissage, et de classement lors de la reconnaissance, les grands avantages des réseaux de neurones résident dans leur capacité **d'apprentissage automatique** (approximation universelle (**Cybenko**, **Hornik**)), ce qui permet de résoudre des problèmes sans nécessiter l'écriture de règles complexes, tout en étant tolérant aux erreurs. Ils résident aussi dans leur capacité à prendre une décision à partir de critères **non formalisables** [14].

Dans ce chapitre, nous présentons donc la conception d'un système de reconnaissance automatique de visage basé sur CNN (Convolutional Neural Network),

4.1 Conception de la méthode proposée

Pour garantir la reproductibilité, l'ensemble d'apprentissage disponible publiquement CASIA2D est utilisé. Nous utilisons aussi pour la validation du système de reconnaissance de visage étudié un échantillon d'images en milieux incontrôlés.

4.1.1 Présentation des bases de données usuelles

4.1.1.1 BDD CASIA 2D milieux contrôlés

La base de données d'images de visage 2DCASIA version 4.0 (ou 2DCASIAV4) utilisée contient 123 images faciales couleur. Toutes les images de visage sont des fichiers BMP couleur de 16 bits et de résolution 640 * 480. Les variations intra-classes typiques incluent illumination, pose, expression, lunettes, distance d'imagerie, etc. [biometrics.idealtest.org/]

Nous avons utilisé 12 images pour chaque personne dans un protocole avec un nombre total de 1476 images ou :

Les images 2, 3, 4, 6, 7, 8, 10, 11, 12 représentent les images de test.

Les images 1, 5, 9 représentent les images d'apprentissage.

Fig.4. 1 *Echantillon d'images BDD CASIA2DV4*

4.1.1.2 BDD milieux incontrôlés

La BDD en milieux incontrôlés que nous avons utilisé pour valider notre SRV, contient 70 personnes collectées du Net avec pour chaque personne 12 images avec différentes variantes. Il s'agit en majorité de personnages célèbres et de footballeurs. Les images sont en 2D, en couleur et avec des tailles différentes. Donc, la première étape de notre travail est le redimensionnement de toutes les images. Un échantillon de la base usuelle est présente dans la figure **Figure 4.2**

Fig.4. 2 *Echantillon d'images BDD en milieux incontrôlés*

Avant de soumettre les images au CNN pour la reconnaissance de visage, nous avons utilisé le même protocole que pour la BDD CASIA2DV4.

4.1.2 Système de reconnaissance automatique de visage basé sur le Deep Learning

Notre système de reconnaissance automatique de visage contient deux phases, la phase d'apprentissage et la phase de test. Pendant la phase d'apprentissage (off-line), l'enrôlement des images faciales est réalisé et les données d'apprentissage sont préparées pour le classificateur. La phase off-line est effectuée une seule fois par le SRV. Dans la phase de test (on-line) les nouvelles données de test sont comparées aux données d'apprentissage. Les étapes de traitement dans les deux phases sont les mêmes. Comme nous l'avons précisé dans le chapitre 3 ; la structure de CNN contient des couches Convolutional, Pooling, Rectified Linear Unit (ReLU) et Fully Connected .

La figure 4.1 illustre les différentes étapes :

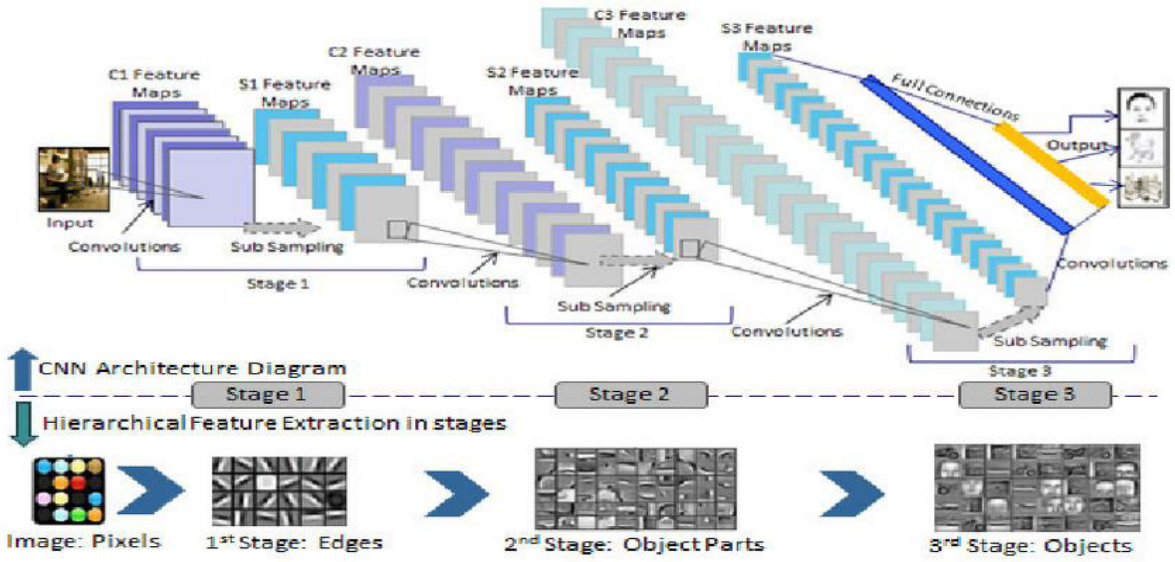


Fig.4. 3 *Echantillon d'images BDD en milieux incontrôlés*

Première étape pour la conception du CNN :

Avant tout, nous devons définir l'architecture du réseau CNN utilisé en construisant les différentes couches de traitement :

1. *Couche de convolution (CONV) qui traite les données d'un champ réceptif ; La couche de convolution est le bloc de construction de base d'un CNN. Le détail de son fonctionnement est précisé dans le chapitre précédent.*

2. *Couche de pooling (POOL), qui permet de compresser l'information en réduisant la taille de l'image intermédiaire (souvent par sous-échantillonnage).*
3. *Couche de correction (ReLU), souvent appelée par abus « ReLU » en référence à la* fonction d'activation *(Unité de rectification linéaire)*
4. *Couche « entièrement connectée » (FC), qui est une couche de type* perceptron *; FC est considéré comme la dernière couche de pool alimentant les fonctionnalités d'un classificateur qui utilise la fonction d'activation Softmax. La somme des probabilités de sortie de la couche entièrement connectée est 1. Ceci est assuré en utilisant le Softmax comme fonction d'activation.*
   La fonction Softmax prend un vecteur de scores réels arbitraires et l'écrase sur un vecteur de valeurs entre zéro et un qui somme à un.

Deuxième étape pour la conception du CNN :

Dans cette étape, nous procédons au paramétrage des couches utilisées : trois hyper paramètres permettent de dimensionner le volume de la couche de convolution (aussi appelé volume de sortie) : la profondeur, le pas et la marge.

- *Profondeur de la couche : nombre de noyaux de convolution (ou nombre de neurones associés à un même champ réceptif).*
- *Le pas contrôle le chevauchement des champs réceptifs. Plus le pas est petit, plus les champs réceptifs se chevauchent et plus le volume de sortie sera grand.*
- *La marge (à 0) ou zero padding : parfois, il est commode de mettre des zéros à la frontière du volume d'entrée.*

4.1.3 *Implémentation et résultats*

Cette section décrit les expériences réalisées et les résultats expérimentaux du système de reconnaissance. Elle contient deux parties pour examiner la performance de l'information dans le système de reconnaissance faciale. Les bases de données utilisées contiennent certains défis et différences. Cela nous permet de tester la technique étudiée et ses algorithmes pour résoudre ces difficultés et problèmes.

Dans la ***première expérience***, le système est testé en utilisant une base de données en milieux incontrôlés composée de seulement 5 personnes pour mettre en œuvre l'algorithme CNN et vérifier sa fonctionnalité et efficacité.

Puis, une fois le CNN opérationnel, nous lançons une ***deuxième expérience*** en testant le SRV basé CNN conçu sur une base de données plus grande. Cette dernière contient 50 personnes

avec 12 images par personne. Les images sont collectées du Net et appartiennent à des célébrités (joueurs de foot, acteurs, chanteurs etc…).

Une ***troisième et dernière expérience*** dans les milieux incontrôlés est menée avec 70 personnes. Notons au passage que nous avons essayé de manipuler la BDD LFW, mais nous avons rencontré des difficultés vu le protocole que nous avons choisi au départ. Rappelons que la LFW BDD présente un nombre d'image variable par personne ce qui représentait un majeur obstacle pour nous compte du temps attribué au projet.

Une ***autre série d'expériences*** est réalisée en utilisant la base de données universelle CASIA2DV4 pour évaluer la stabilité et la performance de notre système en présence des variantes mentionnées ci- dessus.

La section suivante décrit l'algorithme proposé et les résultats expérimentaux obtenus par le système de reconnaissance. Il y a deux parties dans cette étude afin d'examiner la performance de l'information dans le système de reconnaissance faciale. Les bases de données choisies contiennent de nombreux défis et différences. Cela nous permet de tester de nombreuses expériences sur le CNN pour résoudre ces nombreuses difficultés et problèmes.

4.1.3.1 L'algorithme proposé

Le schéma de bloc de l'algorithme de reconnaissance CNN proposé est donné par la **Fig. 4.4** L'algorithme est principalement réalisé en trois étapes comme ci-dessous :

1) Redimensionner les images d'entrée au format ***32x32x3*** ;

2) Construire une structure CNN avec ***trois couches*** constituées respectivement de : filtre convolutif, batch normalisation, ReLu et de regroupement maximum (Max Pooling). Cette couche est répétée trois fois.

3) Après avoir extrait toutes les fonctionnalités, utiliser le classificateur Softmax pour la classification.

Pour sa mise en œuvre le CNN fait appel à la fonction *imageDatastore.* Les dernières versions du MATLAB fournissent la fonction *imageDatastore.* Cette fonction assure:

1. Lecture automatique de lots (batchs) d'images pour un traitement plus rapide dans les applications d'apprentissage automatique et de vision par ordinateur

2. Importation des données à partir de collections d'images trop volumineuses pour tenir dans la mémoire
3. Étiquetage des données d'image automatiquement en fonction des noms de dossier
4. Apprentissage plus en profondeur en utilisant l'apprentissage par transfert

Seulement, au début de notre projet nous avons rencontré des difficultés dans l'utilisation de la fonction *imageDatastore.*

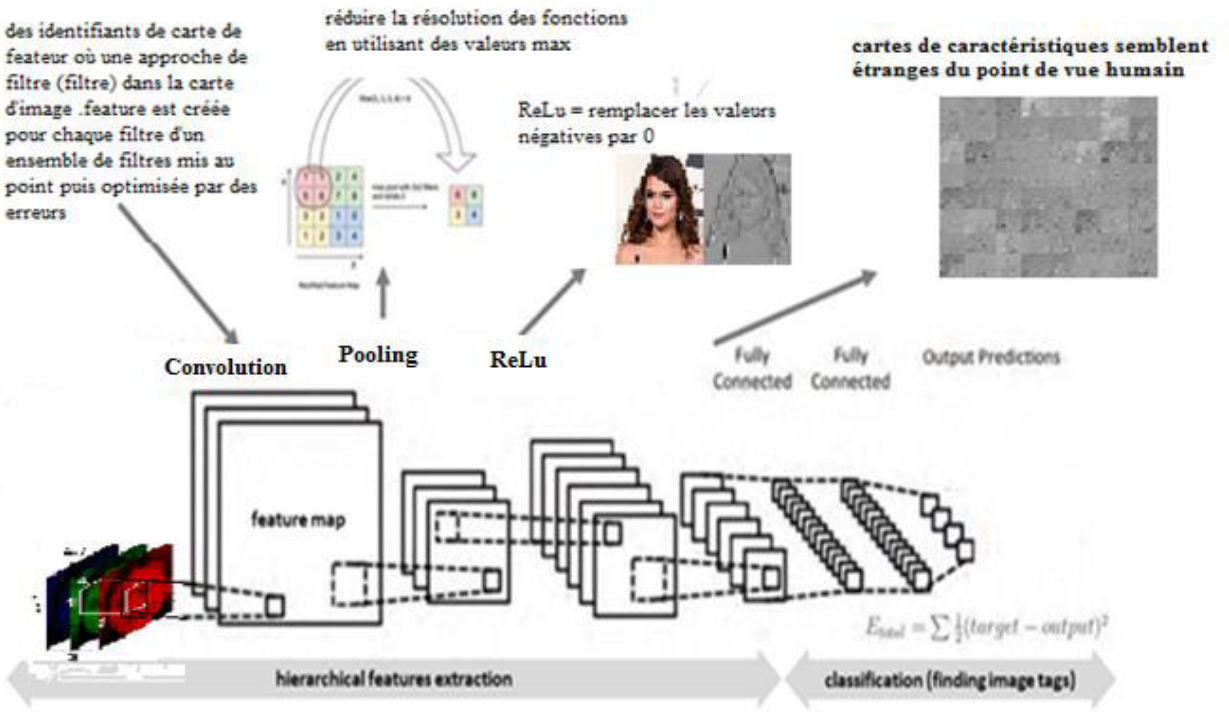


Fig.4. 4 *Schéma de bloc de l'algorithme proposé*

Ces difficultés sont dues à la version du MATLAB2015, pour surmonter ces problèmes nous avons conçus une fonction nous assurant :

- Création de dossiers R, G, B
- Création des sous dossiers « personnes » contenants le nombre d'images par personne
- Indexage des personnes
- Indexage des images par personne.

Par la suite nous avons réalisé une détection des images de la BDD utilisée par le détecteur de Viola Jones voir figure 4.5.

Fig.4. 5 Exemple de détection par Viola Jones des images Sur la BDD CASIAV4

Ensuite, l'image est redimensionnée à la taille 32x32 pour être prise en compte par le CNN voir figure 4.6.

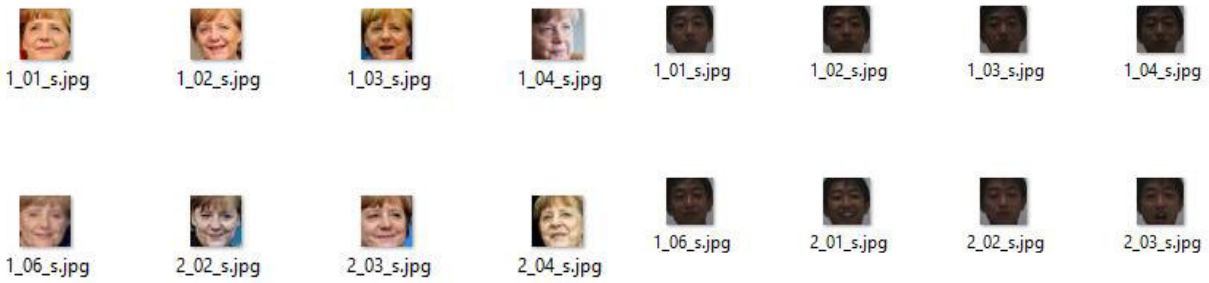


Fig.4. 6 Exemple d'image redimensionnée

Après la détection des images et leur redimensionnement, nous présentons les trois composantes couleurs R,V,B au réseau CNN pour la reconnaissance.

4.1.3.2 Définition de l'architecture et propriétés du CNN utilisé

Pour la couche d'entrée la fonction *imageInputLayer* est utilisée.

1 Création d'une couche d'entrée d'image :

Layer = imageInputLayer(inputSize,Name,Value)

Cette fonction définit les propriétés facultatives à l'aide des arguments de la paire (nom,valeur). On peut spécifier plusieurs paires (nom,valeur). Chaque nom de propriété est placé entre guillemets simples. Nous créons une couche d'entrée d'image pour les images couleur 32 x 32 avec le nom «entrée». Par défaut, la couche effectue la normalisation des données en soustrayant l'image moyenne du jeu d'apprentissage de chaque image d'entrée.

```
%% Define the CNN and Train it using Training Data
layers = [  imageInputLayer( [ Height, Width, Channel ] )
```

```
ImageInputLayer with properties:

              Name: ''
         InputSize: [32 32 3]

 Hyperparameters
  DataAugmentation: 'none'
     Normalization: 'zerocenter'
```

Propriétés :

1. *InputSize* (taille de l'entrée) : La taille des données d'entrée est définie comme un vecteur ligne avec trois valeurs [32 32 3], où h est la hauteur, w est la largeur et c est le nombre de canaux.
2. *DataAugmentation* : transformation d'augmentation de données. La transformation d'augmentation de données à utiliser pendant l'entraînement est spécifiée dans notre cas par '*none*' - Aucune augmentation de données
3. *Normalisation* - Transformation de données

   La transformation de données à appliquer chaque fois que les données sont propagées vers l'avant à travers la couche d'entrée, spécifiée comme l'une des suivantes : '*zerocenter*' - La couche soustrait l'image moyenne de l'ensemble d'apprentissage. '*none*' - Pas de transformation.
4. *Nom* - Nom du Layer : Nom de la couche, spécifié en tant que vecteur de caractères. Si *Nom* est réglé sur '', le logiciel attribue automatiquement un nom au moment de la formation.

2 Création de la convolution :

La couche convolution est crée par la fonction ***convolution 2dLayer***

layer = convolution2dLayer(filterSize,numFilters,Name,Value)

Arguments d'entrée : Utiliser des paires d'arguments (nom,valeur) séparées par des virgules pour spécifier la taille du remplissage de zéros à ajouter le long des bords de l'entrée de couche ou pour définir les propriétés *Stride*, *NumChannels*, *WeightLearnRateFactor*, *BiasLearnRateFactor*, *WeightL2Factor*, *BiasL2Factor* et Name.

- *Nom - Nom de couche*
- *FilterSize - Hauteur et largeur des filtres*
- *NumFilters - Nombre de filtres*
- *Stride - Taille de l'étape pour l'entrée de déplacement*
- *PaddingSize - Taille du rembourrage*
- *PaddingMode - Méthode pour déterminer la taille de remplissage*
- *Rembourrage - Taille du rembourrage*
- *Poids - Poids des couches*
- *NumChannels - Nombre de canaux pour chaque filtre*

```
%% Define the CNN and Train it using Training Data
layers = [  imageInputLayer( [ Height, Width, Channel ] )

            convolution2dLayer( 6, 64, 'Padding', 'same' )
```

```
layers =

  15x1 Layer array with layers:

     1   ''   Image Input             32x32x3 images with 'zerocenter' normalization
     2   ''   Convolution             64 8x8 convolutions with stride [1  1] and padding 'same'
     3   ''   Batch Normalization     Batch normalization
     4   ''   ReLU                    ReLU
     5   ''   Max Pooling             2x2 max pooling with stride [2  2] and padding [0  0  0  0]
     6   ''   Convolution             32 8x8 convolutions with stride [1  1] and padding 'same'
     7   ''   Batch Normalization     Batch normalization
     8   ''   ReLU                    ReLU
     9   ''   Max Pooling             2x2 max pooling with stride [2  2] and padding [0  0  0  0]
    10   ''   Convolution             16 8x8 convolutions with stride [1  1] and padding 'same'
    11   ''   Batch Normalization     Batch normalization
    12   ''   ReLU                    ReLU
    13   ''   Fully Connected         1 fully connected layer
    14   ''   Softmax                 softmax
    15   ''   Classification Output   crossentropyex
```

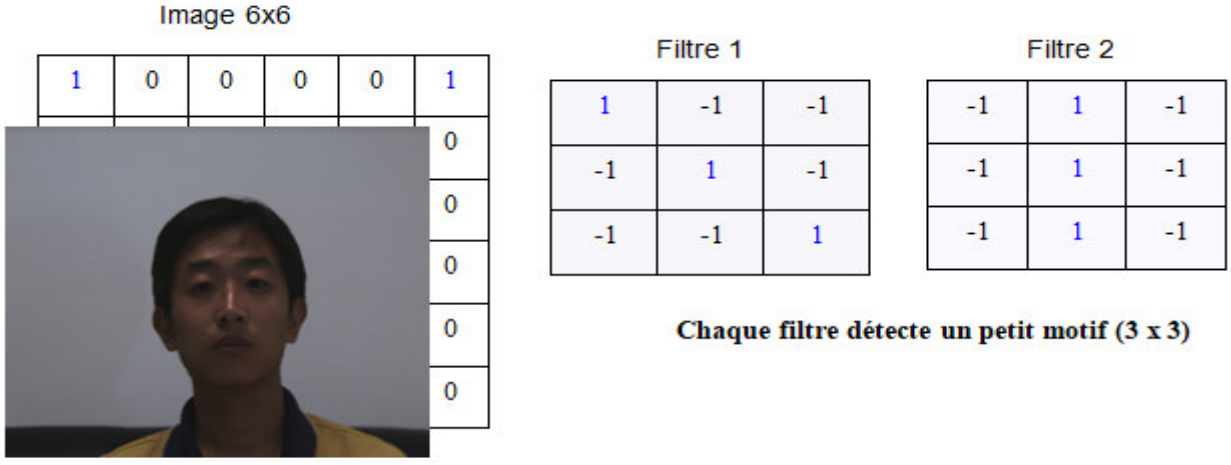


Fig.4. 7 exemple d une image 6x6 convolu

3 Création de la normalisation :
batchNormalizationLayer

Une couche de normalisation par lots normalise chaque canal d'entrée sur un mini-lot. La couche normalise d'abord les activations de chaque canal en soustrayant la moyenne du mini-lot et en divisant par l'écart-type du mini-lot. Ensuite, la couche décale l'entrée d'un décalage appréhensible β et la met à l'échelle par un facteur d'échelle approchant γ. Utilisez des couches de normalisation par lots entre les couches convolutives et les non-linéarités, telles que les couches ReLU, pour accélérer l'apprentissage des réseaux de neurones convolutionnels et réduire la sensibilité à l'initialisation du réseau.

Propriétés générales de la couche

Nom - Nom de la couche

Num Channels - Nombre de canaux d'entrée

Epsilon - Constante à ajouter aux variances mini-batch

```
%% Define the CNN and Train it using Training Data
layers = [  imageInputLayer( [ Height, Width, Channel ] )

            convolution2dLayer( 6, 64, 'Padding', 'same' )


            batchNormalizationLayer
```

4.1.3.3 Expérimentation et résultats

1. Série d'expérience sur BDD en milieux incontrôlés

1. a) ***Première expérience :*** Cas de 20 personnes NG (BDD constituée d'acteurs célèbres images en NG, size 640x680, format JPG se présentant de la même façon et protocole que les autres BDD utilisées dans nos expériences).

L'échantillon de quelques images est donné par la figure 4.8

Fig.4. 8 L'échantillon de quelques images utilisé

L'architecture du réseau est confirmée par la figure 4.9 suivante :

```
layers =

  8x1 Layer array with layers:

     1   ''   Image Input             32x32x3 images with 'zerocenter' normalization
     2   ''   Convolution             64 6x6 convolutions with stride [1  1] and padding 'same'
     3   ''   Batch Normalization     Batch normalization
     4   ''   ReLU                    ReLU
     5   ''   Max Pooling             2x2 max pooling with stride [2  2] and padding [0  0  0  0]
     6   ''   Fully Connected         1 fully connected layer
     7   ''   Softmax                 softmax
     8   ''   Classification Output   crossentropyex
```

Fig.4. 9 L'architecture du réseau Proposé

Les résultats obtenus dans la première expérience confirment bien la fonctionnalité du réseau CNN pour la reconnaissance. Les taux de performance du SRV CNN sont donnés par les courbes représentées sur la figure suivant.

Definition Accuracy :

Accuracy = sum (TestPredLabel == TestLabels) / numel(TestLabels)

Accuracy: 0.85

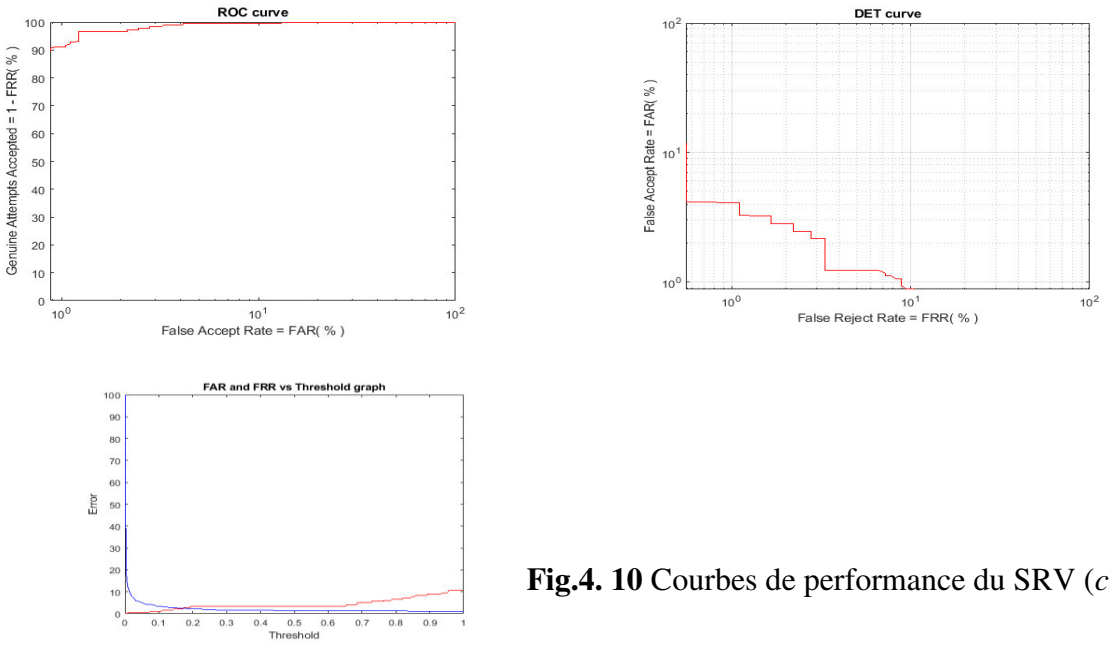
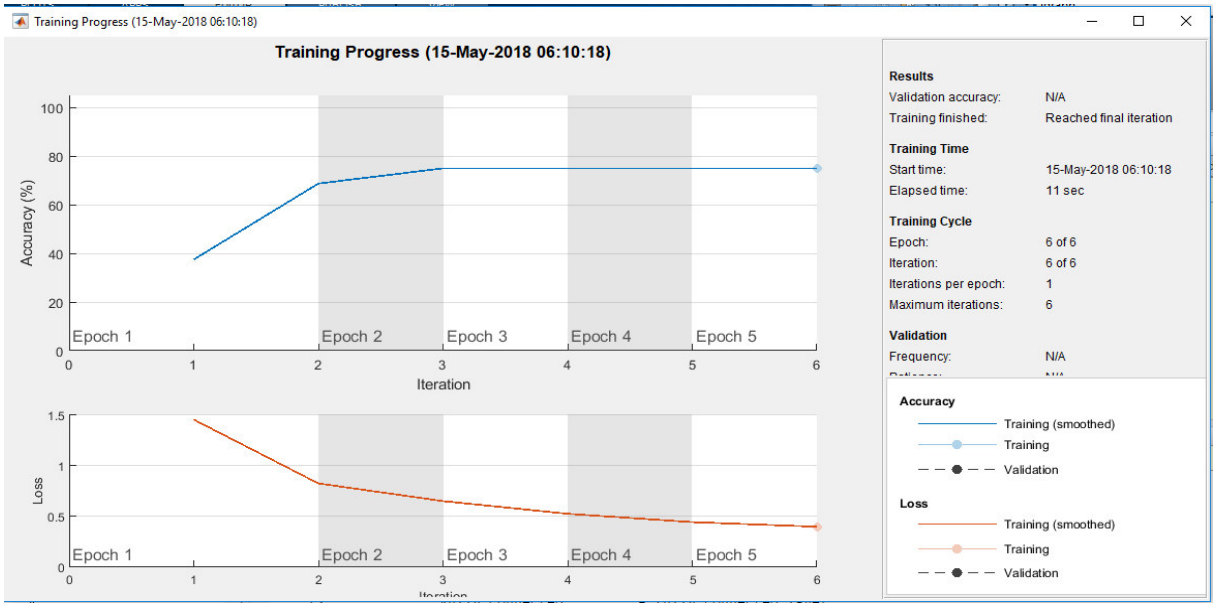


Fig.4. 10 Courbes de performance du SRV (*cas de 20 Person* N&G)

Discussion 1 : Les performances obtenus sont relativement faible (85%) compte tenu la taille réduite de l'échantillon d'images de la BDD et le fait qu'il s'agit d'images NG. Ceci s'explique par le fait que le CNN est plus performant dans le cas de BDD de grandes tailles. Une question se pose est ce que la couleur des images est une information pertinente pour le CNN ? Pour répondre à cette question, nous avons entamé une autre expérience avec 20 mêmes personnes RGB.

1.b) ***Deuxième expérience*** : cas de 20 Personnes en RGB. Il s'agit des mêmes personnes. Seulement, dans la programmation du réseau, nous changeons le paramètre c correspondant au nombre de canaux.

La phase d'apprentissage du CNN présente des caractéristiques selon ce qui suit :



Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:03	5.00%	3.0336	0.0100
30	30	00:01:42	95.00%	0.0745	0.0100

Les performances du SRV CNN sont montrées dans les figures 4.11

(Distribution des clients et imposteurs dans la BDD)

Accuracy: 0.994

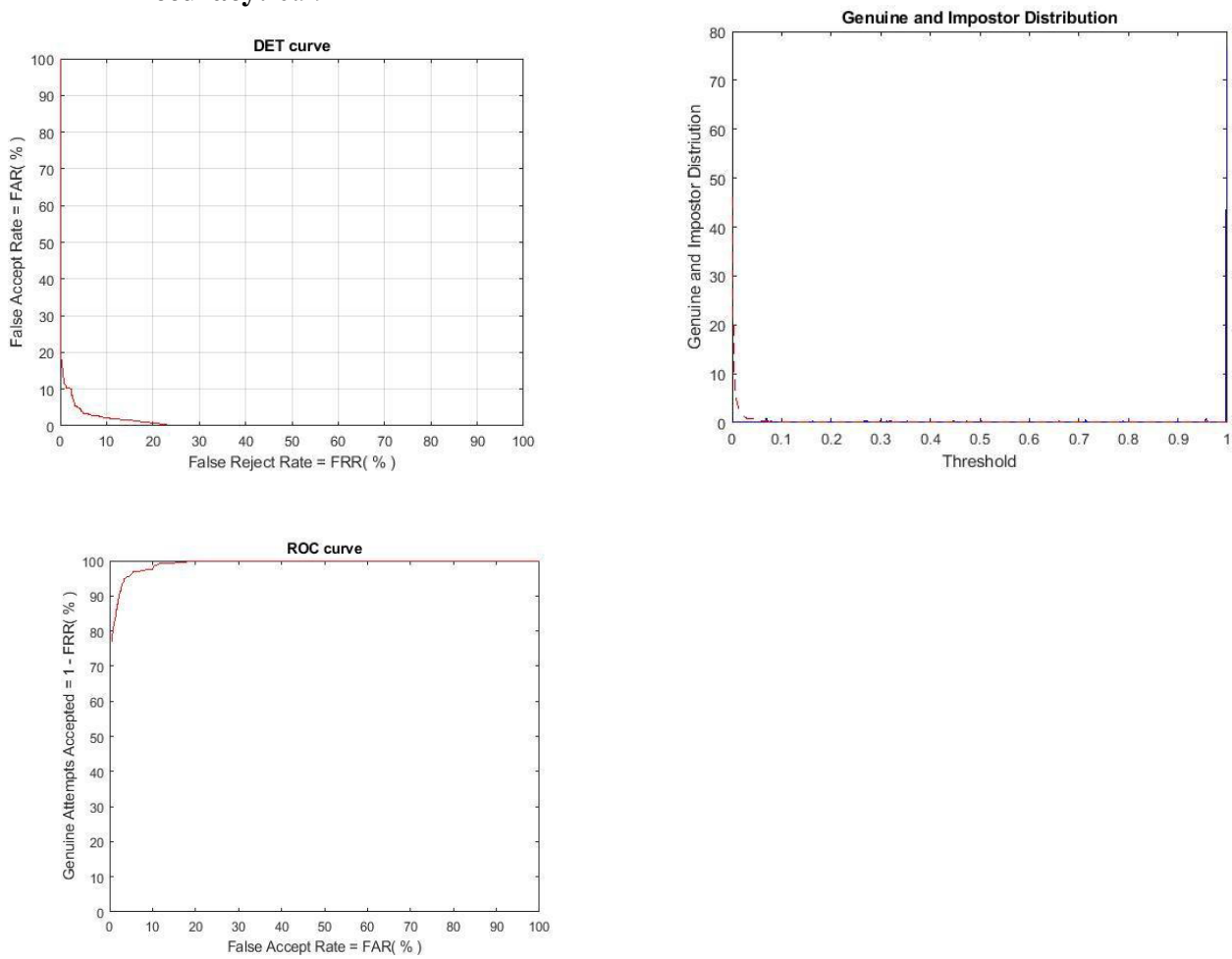


Fig.4. 11 *Courbes de performance du SRV (cas de 20 Person R V B)*

Discussion 2 : Nous constatons à travers cette deuxième expérience que la couleur est une information très importante pour le réseau CNN. Cela s'explique par le fait que, dans le cas des images couleur, la première couche du réseau CNN (couche d'entrée (input layer)) reçoit les trois matrices des composantes couleur R, V, B (voir figure 4.4) . celles-ci sont traitées par toutes les couches du réseau séparément. Ce qui aide l'apprentissage à extraire l'information caractéristique pertinente.

2. Deuxième série d'expériences en milieu contrôlé (BDD CASIA2DV4)

2.a) 1ère expérience : cas de 50 personnes : dans cette expérience, nous choisissons 50 personnes (0001_0050) de la BDD CASIA2DV4. Nous voulons testé les performance du SRV CNN sur un échantillon réduit d'une BDD présentant différentes variantes (poses, expression, illumination,) en milieu contrôlé.

Accuracy : 0.98

```
%% Define the CNN and Train it using Training Data
layers = [  imageInputLayer( [ Height, Width, Channel ] )

            convolution2dLayer( 6, 64, 'Padding', 'same' )
            batchNormalizationLayer
            reluLayer

            maxPooling2dLayer( 2, 'Stride', 2 )

            fullyConnectedLayer( nPerson )
            softmaxLayer
            classificationLayer ];

options = trainingOptions('sgdm');
%   'MaxEpochs',10, ...
%   'ValidationFrequency',30,...
%   'Verbose',false,...
%   'Plots','training-progress');

convnet = trainNetwork( TrainSet, TrainLabels, layers, options );
```

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:18	0.00%	5.3038	0.0100
13	50	00:04:03	99.22%	0.4057	0.0100
25	100	00:07:36	100.00%	0.0142	0.0100

Fig.4. 12 *Architecteur de réseaux CNN (cas de 50 Person BDD CASIA2DV4)*

2.a) 2ème expérience : cas de 123 personnes : dans cette deuxième expérience, nous validons nos travaux sur la totalité de la BDD CASIA2DV4. Nous voulons ici testé les performance du SRV CNN sur un échantillon plus grand d'une BDD présentant différentes variantes (poses, expression, illumination,) en milieu contrôlé.

Accuracy: 0.994

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:12	0.00%	6.2166	0.0100
5	50	00:06:07	55.47%	4.1537	0.0100
10	100	00:12:08	100.00%	0.5201	0.0100
14	150	00:18:03	100.00%	0.0428	0.0100
19	200	00:24:01	100.00%	0.0200	0.0100
23	250	00:29:57	100.00%	0.0149	0.0100
28	300	00:35:55	100.00%	0.0115	0.0100
30	330	00:39:26	100.00%	0.0104	0.0100

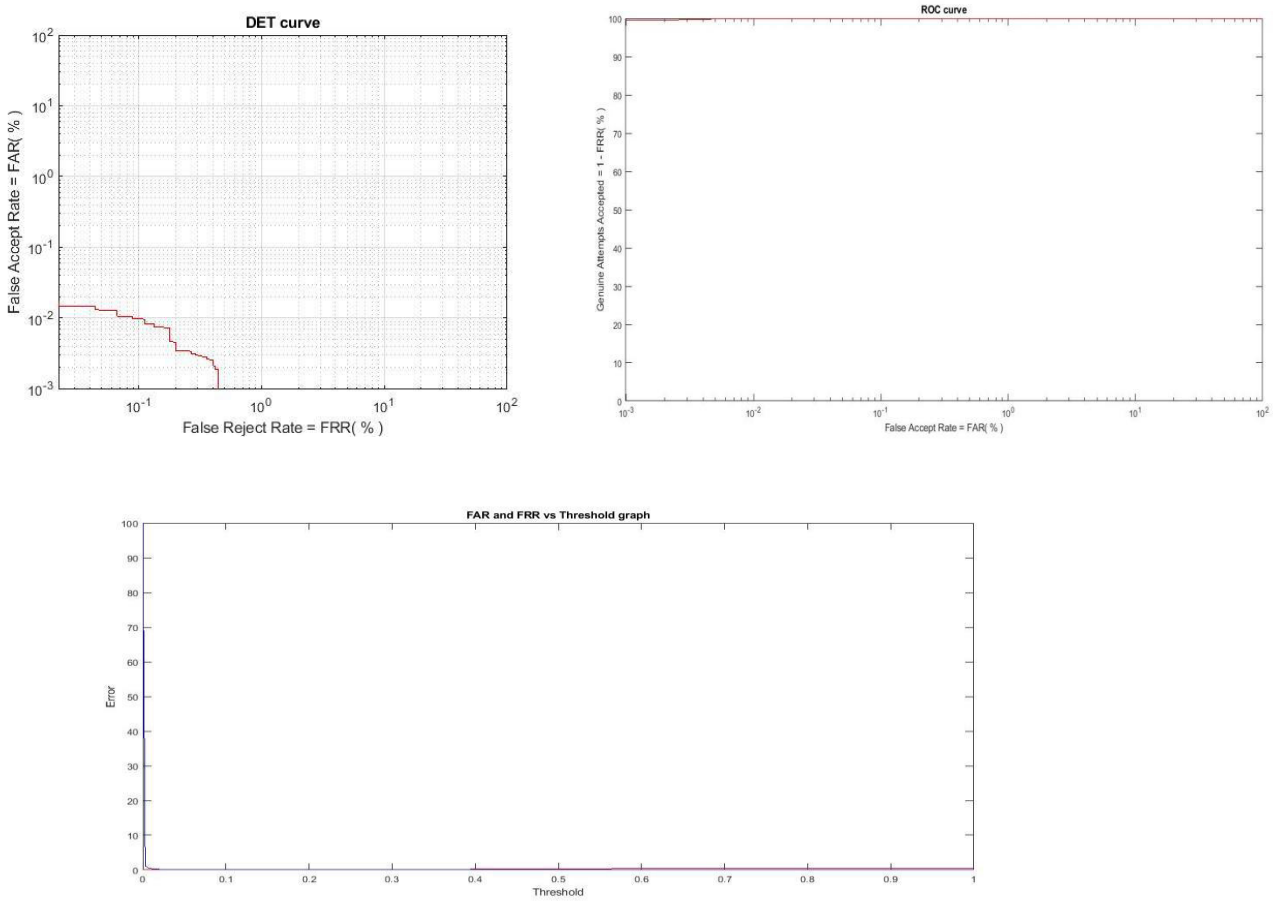


Fig.4. 13 *Courbes de performance du SRV (cas de 123 Person BDD CASIA2DV4)*

Discussion : Dans cette expérience les performances obtenues montrent que malgré l'échantillon réduit de la BDD, le CNN reste très efficace. Ces résultats obtenus sont meilleurs que ceux obtenus dans e cas des expériences menés sur la BDD en milieux incontrôlés. Ce qui est évident car dans le cas des environnements incontrôlés les variantes

sont multiples et nécessite un apprentissage plus approfondi ou un jeu de données plus grand pour l'apprentissage.

Visualizing Activation in GoogLeNet

```
net = googlenet
```

```
net =
  DAGNetwork with properties:

         Layers: [144×1 nnet.cnn.layer.Layer]
    Connections: [170×2 table]
```

Regardons seulement les premières couches :

```
net.Layers(1:5)
```

```
ans =
  5x1 Layer array with layers:

     1   'data'             Image Input                   64x64x3 images
with 'zerocenter' normalization
     2   'conv1-7x7_s2'     Convolution                   64 7x7x3
convolutions with stride [2  2] and padding [3  3  3  3]
     3   'conv1-relu_7x7'   ReLU                          ReLU
     4   'pool1-3x3_s2'     Max Pooling                   3x3 max pooling
with stride [2  2] and padding [0  1  0  1]
     5   'pool1-norm1'      Cross Channel Normalization   cross channel
normalization with 5 channels per element
```

Afficher toutes les propriétés

Les hyper paramètres nous disent que cette couche effectue 64 opérations de filtrage différentes sur les canaux d'entrée, et chaque filtre est 7x7x3. La valeur de la foulée [2 2] nous indique que la sortie du filtre est sous-échantillonnée d'un facteur 2 dans chaque direction.

Pour expérimenter avec ce réseau, je vais utiliser une image que j'ai prise de moi-même tout à l'heure. Je vais aller de l'avant et le redimensionner à la taille attendue par le réseau.

```
im = imread('hocin.jpg');
im = imresize(im,net.Layers(1).InputSize(1:2));
imshow(im)
```

Fig. 4.14 Image d'entrée (64x64x3)

Utilisez la fonction d'activation pour calculer les activations neuronales à partir de la couche conv1-7x7_s2.

J'interprète la taille de l'acte en disant que la sortie de cette couche comprend 64 images 112x112 différentes. La plage de valeurs d'activation est comprise entre -3 000 et 3 000.

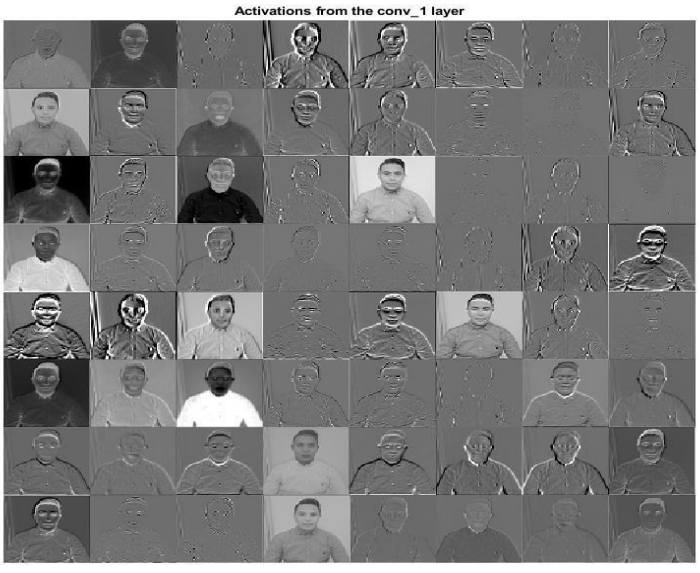
```
act = activations(net,im,'conv1-7x7_s2','OutputAs','channels');
size(act)
act = reshape(act,size(act,1),size(act,2),1,size(act,3));
act_scaled = mat2gray(act);
montage(act_scaled)
```

Les fonctions mat2gray et montage sont utiles pour redimensionner ces images à la plage [0,1] et ensuite les afficher ensemble.

Fig. 4. 16 montage la couche conv1-7x7_s2.

Je pense qu'il sera plus facile de voir et de comparer les images d'activation individuelles si nous appliquons un étirement de contraste. J'utiliserai les fonctions imadjust et stretchlim de l'Image Processing Toolbox. (Il y a un peu de code supplémentaire pour gérer le fait que stretchlim et imadjust ne supportent pas les entrées multidimensionnelles.)

```
tmp = act_scaled(:);
tmp = imadjust(tmp,stretchlim(tmp));
act_stretched = reshape(tmp,size(act_scaled));
montage(act_stretched)
title('Activations from the conv_1layer','Interpreter','none')
```



C'est un peu trop de moi à la fois. Zoom sur quelques images d'activation

```
subplot(1,2,1)
imshow(act_stretched(:,:,:,33))
title('Channel 33')
subplot(1,2,2)
imshow(act_stretched(:,:,:,34))
title('Channel 34')
```

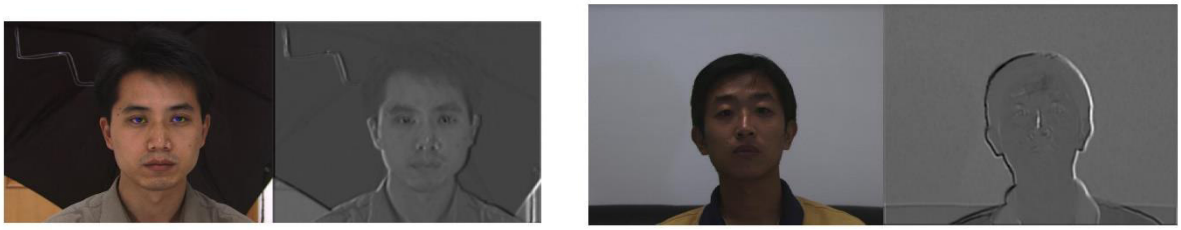


Fig. 4. 17 des images de composant de gradient

Regardons juste une autre couche, celle qui suit immédiatement.

C'est une couche d'unité linéaire rectifiée. Une telle couche ne fait que clipser tout nombre négatif à 0. Cela signifie que toute la variation des valeurs négatives de la sortie de la couche précédente est supprimée. A quoi cela ressemble-t-il?

```
act2 = activations(net,im,'conv1-relu_7x7');
```

```
act2 = reshape(act2,size(act2,1),size(act2,2),1,size(act2,3));
act2_scaled = mat2gray(act2);
tmp = act2_scaled(:);
lim = stretchlim(tmp);
lim(1) = 0;
tmp = imadjust(tmp,lim);
act2_stretched = reshape(tmp,size(act2_scaled));
clf
montage(act2_stretched)
title('Activations from the conv1-relu_7x7 layer','Interpreter','none')
```

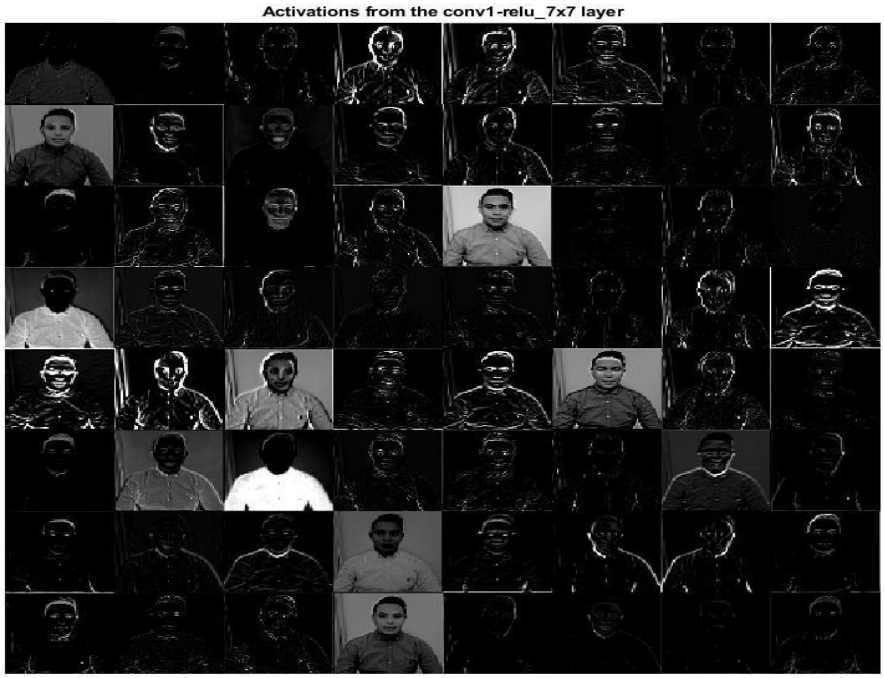


Fig. 4. 18 des images de composant de gradient

Conclusion

Nous avons présenté dans ce chapitre l'implémentation de l'approche de reconnaisance de visage basée sur les réseaux de neurones convolutionnels, pour cela nous avons avons utilisé deux modèles d'architectures et plusieurs expériences et présenté différents résultats obtenus

en termes de précision et d'erreur. La comparaison des résultats trouvés a montré que le nombre d'époque, la taille de la base et la profondeur de réseaux, sont des facteurs importants pour l'obtention de meilleurs résultats.

L'approche proposée pour la reconnaissance faciale fonctionne pour de basses résolutions (les images sont de taille 32 × 32), intéressantes dans le cadre de futures extensions de la technique. L'approche utilise une architecture particulière de réseau de neurones convolutionnels. Celle-ci projette un visage dans un espace de plus faible dimension, où la reconnaissance proprement dite est effectuée. Les tests effectués sur une base restreinte étant encourageants,

Conclusion générale

Au cours des trois dernières années, principalement en raison des progrès de l'apprentissage en profondeur, plus concrètement des réseaux convolutifs, la qualité de la reconnaissance d'images et de la détection d'objets a progressé à un rythme spectaculaire. Une des nouvelles encourageantes est que la plupart de ces progrès ne sont pas seulement le résultat d'un matériel plus puissant, de jeux de données plus volumineux et de modèles plus grands, mais principalement une conséquence de nouvelles idées, d'algorithmes et d'architectures réseau améliorées. Aucune nouvelle source de données n'a été utilisée, par exemple, par les meilleures contributions au concours 2014 de l'ILSVRC (ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC2014) en plus de l'ensemble de données de classification du même concours à des fins de détection et reconnaissance. Ceci nous motive et nous encourage à nous lancer dans cette voie de l'apprentissage en profondeur afin de surmonter les défis : 1) la malédiction de la dimensionalité, 2) la rapidité du processus de reconnaissance de visage, 3) la robustesse du SRV en milieux incontrôlés, 4) la taille réduite des BDD pour l'apprentissage des SRV et d'autres…

Cette dernière partie de conclusion dresse un bilan des travaux effectués dans le cadre de ce projet de fin d'étude. Nous rappellerons dans un premier temps les principales contributions de ces travaux. Nous discuterons ensuite les limitations des modèles étudiés, ainsi que les améliorations potentielles à apporter. Nous présenterons ensuite quelques pistes de travaux futurs et des perspectives de recherche.

1 Récapitulatif des contributions

Quand nous avons dû faire face au problème, la première étape a été de faire des recherches sur l'histoire et l'état actuel du terrain, comme expliqué au chapitre 2. Cela nous a donné une bonne base sur ce à quoi nous pouvions nous attendre et quelles méthodes éviter ? Tout d'abord, nous avons brièvement envisagé d'utiliser l'approche CNN, en utilisant certains des logiciels existants pré-entraîné afin de réaliser l'apprentissage en profondeur des caractéristiques faciales pertinentes, et de les utiliser par la suite comme des fonctionnalités dans un classificateur. Cela nous a séduits par son intuitivité:

le processus pouvait être facilement compris, quels que soient les détails mathématiques spécifiques.

Cependant, certains obstacles liés à la version du Matlab au départ, l'indisponibilité de BDD visage dans les réseaux pré-entrainés et les recherches existantes nous mènent, de manière assez prévisible, au domaine des CNN.

Nous nous sommes intéressés dans ce mémoire à la problématique de la reconnaissance automatique de visage. L'idée était de se démarquer de la méthodologie dominante qui se base sur l'utilisation de caractéristiques conçues manuellement, en proposant des modèles qui soient les plus génériques possibles et indépendants du domaine. Ceci a été réalisé en automatisant la phase d'extraction des caractéristiques, qui sont dans notre cas générées par apprentissage à partir d'exemples, sans aucune connaissance a priori. Les contributions de ce travail concernent donc principalement la phase d'apprentissage des caractéristiques et la résolution du problème de dimensionnalité.
Ce travail s'intéresse à la problématique de la reconnaissance automatique des personnes. L'idée est de se démarquer de la méthodologie dominante qui se base sur l'utilisation de caractéristiques conçues manuellement, et de proposer des modèles qui soient les plus génériques possibles et indépendants du domaine. Ceci est fait en automatisant la phase d'extraction des caractéristiques, qui sont dans notre cas générées par apprentissage à partir d'exemples, sans aucune connaissance a priori. Nous nous appuyons pour ce faire sur des travaux existants sur les modèles neuronaux pour la reconnaissance d'objets dans les images fixes, et nous étudions leur application à la reconnaissance de visage. Plus concrètement, nous proposons deux modèles d'apprentissage profond des caractéristiques pour la reconnaissance :
- Un modèle d'apprentissage profond avec création du réseau CNN, qui peut être vu comme la mise en œuvre du réseau CNN. - Un modèle d'apprentissage profond, qui se base sur un réseau de transfert, utilisant un réseau pré-entraîné GoogleNet.

2 Justification du choix de la méthodologie

Comme nous l'avons déjà expliqué, l'apprentissage en profondeur fournit de nouvelles références dans de nombreuses applications de la vision par ordinateur. De plus, en raison de l'intérêt pour le domaine, il y a beaucoup de recherches. Plus encore, même si certaines de ces

recherches appartiennent à des entreprises et sont donc privées, de nombreux articles sont accessibles au public. Ces deux facteurs - qualité des résultats et disponibilité de l'information - nous amènent à choisir d'utiliser les CNN dans notre système de reconnaissance faciale.

Notre choix s'est donc porté sur la méthodologie CNN, il est fondé sur les faits suivants :

1. CNN sont une famille de réseaux neuronaux multicouches particulièrement conçus pour une utilisation sur des données bidimensionnelles, telles que des images et des vidéos.
2. Les CNN sont influencés par des travaux antérieurs dans les réseaux neuronaux temporisés (TDNN), qui réduisent les besoins de calcul d'apprentissage en partageant des poids dans une dimension temporelle et sont destinés au traitement du signal et de l'image.
3. CNN sont les premiers vraiment approche réussie d'apprentissage en profondeur où de nombreuses couches d'une hiérarchie sont formées avec succès de manière robuste.
4. Un CNN est un choix de topologie ou d'architecture qui exploite les relations spatiales pour réduire le nombre de paramètres qui doivent être appris et améliore ainsi la propagation générale inverse d'entraînement. CNN ont été proposés comme un cadre d'apprentissage en profondeur qui est motivé par des exigences minimales de prétraitement des données.
5. Le réseau de neurones convolutionnels (CNN) s'est avéré être une approche très efficace pour reconnaître les visages. Par rapport aux méthodes traditionnelles, leur approche présente l'avantage qu'aucun espace supplémentaire n'est nécessaire pour stocker les caractéristiques faciales puisque les images sont recadrées en ligne.
6. De plus, grâce à la formation de bout en bout, cette approche fait un meilleur usage des interactions entre les caractéristiques locales dans le modèle. Deux réseaux CNN de référence (c'est-à-dire AlexNet et ResNet) sont utilisés pour analyser l'efficacité des méthodes.

Les expériences montrent que le système proposé atteint des performances comparables avec d'autres méthodes de pointe sur les visages étiquetés dans la nature et les tâches de vérification des visages sur YouTube.

Pour toutes ces raisons, notre motivation est grande pour faire des investigations dans cette piste afin de résoudre le problème de la dimensionnalité qui s'accroît de jour en jour avec la technologie.

3. Récapitulatif de l'expérimentation et résultats

Cette étude a été réalisée en se basant sur le réseau CNN adapté à la problématique de la reconnaissance de visage, puisque ce type de réseau a prouvé son succès dans le domaine de la détection d'objets, la classification et l'étiquetage. Ceci nous a permis de réaliser la reconnaissance automatique et à identifier le modèle de réseau CNN le plus performant (un réseau de neurones convolutionnel à quinze couches), et de justifier son utilisation pour le reste des expérimentations.

Enfin, afin de valider la généricité des deux modèles proposés, ceux-ci ont été évalués sur deux problématiques différentes, à savoir la reconnaissance de visage en milieux incontrôlés avec une BDD à taille réduite (sur la base collectée du web : footballeurs et célébrité), et la reconnaissance faciales en milieu contrôlé (sur la base CASIA2DV4 (diverses variantes de visages)). L'étude des résultats a permis de valider les approches, et de montrer qu'elles obtiennent des performances acceptables (avec 90% de bonne reconnaissance pour la base en milieux incontrôlés, et 90% pour la base CASIA2DV4).

4. Perspectives

Plusieurs pistes sont envisageables :

Pour l'approche d'auto-encodage concernant la partie pooling ;

Agir au niveau de la couche convolution pour l'amélioration de l'extraction des caractéristiques ;

Enfin, il serait intéressant d'étudier la possibilité de proposer un modèle "hybride", qui combine l'apprentissage. En effet, de nombreux travaux dans le domaine de la reconnaissance d'objets ont démontré l'intérêt de pré-entraîner de manière, les couches inférieures d'un modèle profond, et d'utiliser ce modèle "intermédiaire" pour initialiser un apprentissage supervisé (un procédé communément appelé fine-tuning dans la littérature).

Ainsi, ce mémoire présente un résumé sur l'état actuel du domaine de l'apprentissage en profondeur de la machine et une perspective sur la façon dont il peut évoluer.

Références Bibliographique

[1] John D. Woodward, Jr., Christopher Horn, Julius Gatune, and Aryn Thomas, “Biometrics A Look at Facial Recognition”, documented briefing by RAND Public Safety and Justice for the Virginia State Crime Commission, 2003.

[2] H. Guesmi, "Identification de personnes par fusion de différentes modalités biométriques," Télécom Bretagne; Université de Rennes 1, 2014.

[3] M.Belahcene, “Identification et Authentification en Biométrie ,Thèse de doctorat,’’ Université Mohamed KHIDER, BISKRA, Janvier 2013

[4]R.Yapp, april 2011. Brazilian police to use robocop-style glasses at world cup. The Telegraph. http://www.telegraph.co.uk/news/worldnews/southamerica/brazil/8446088 .

[5] B.Khefif , « Mise au point d’une application de reconnaissance faciale » Université Abou Bakr Belkaid – Tlemcen, 28 novembre 2013

[6] A.Ouamane, " Reconnaissance Biométrique par Fusion Multimodale du Visage 2D et 3D Thèse de doctorat,’’ Université Mohamed KHIDER, BISKRA, Juin 2015

[7] E.Hadjaidji « Modélisation d’empreinte biométrique par un modèle flou de Sugeno optimisé » Master Université Kasdi Merbah– Ouargla | 2017

[8] S. Liu, M. Silveman, « A pratical Guide to Biometric Security Technology », IEEE Computer Society, IT Pro Security, Janvier 2001.

[9] A. K. Jain, L. Hong, S. Pankanti, « Biometrics : Promising Frontiers for Emerging Identification Market », Communications of the ACM, pp. 91-98, February 2000

[10] R. Hietmeyer. Biometric Identification promises fast and secure processing of airline passengers.Internation Civil Aviation Organization Journal, vol. 55, no. 9, pp. 10-11, 2000.

[11] P. Buyssens, Fusion de différents modes de capture pour la reconnaissance du visage appliquée aux e_transactions, PhD thesis l’université de caen spécialité : informatique et applications, 07/07/2006

[12] M.R. Alismail, N.Ourchani., "Fusion multimodale des scores pour la reconnaissancedes personnes", Master 2, Université Mohamed Khider Biskra, 2011.

[13] https://www.mathworks.com/help/nnet/ug/introduction-to-convolutional-neural-networks.html

[14] H. D. Hubel, and T. N. Wiesel, " Receptive Fields of Single neurones in the Cat's Striate Cortex." Journal of Physiology. Vol 148, pp. 574-591, 1959.

[15] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, Massachusetts: The MIT Press, 2012.

[16] P. Omkar, A Vedaldi, and A Zisserman. "Deep face recognition." British Machine Vision Conference. Vol. I. No. 3. 2015

[17] R. Beveridge and M. Kirby. Biometrics and Face Recognition. IS&T Colloquium, pp. 25, 2005.

[18] A. Aissaoui, "Reconnaissance bimodale de visages par fusion de caractéristiques visuelles et de profondeur," Lille 1, 2014.

[19] Ethem Alpaydin. 2010. Introduction to Machine Learning (2nd ed.). The MIT Press https://www.mitpress.mit.edu/books/introduction-machine-learning

[20] https://www.syr-res.com?R7895 September 29, 2015

[21] https://www.mathworks.com/training-schedule/deep-learning onramp.html/

[22] SVM requires Statistics and Machine Learning Toolbox

[23]Introducing Deep Learning with MATLAB

[24] I. Arel, C. Rose, and T. Karnowski. Deep machine learning a new frontier in artificial intelligence. IEEE Computational Intelligence Magazine, 5:13–18, November 2010.

[25] Y. Bengio. Learning deep architectures for AI. in Foundations and Trends in Machine Learning, 2(1):1–127, 2009.

[26] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives.

IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 38:1798–1828, 2013

[27] Voir Wikipédia http://en.wikipedia.org/wiki/Deep_learning on "Deep Learning" à partir de cette dernière mise à jour en octobre 2013

[28] https://www.symmetrymagazine.org/article/deep-learning-takes-on-physics /

[29] http://deeplearning.net/reading-list /

[30] http://ufldl.stanford.edu/wiki/index.php/

[31] http://www.cs.toronto.edu / hinton /

[32] http://deeplearning.net/tutorial /

[33] L. Deng and D. Yu. Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing, vol. 7, nos. 3–4, pp. 197–387, 2013. DOI: 10.1561/2000000039.

[34] The MNIST database of handwritten digits [Online].

[35]http://yann.lecun.com/exdb/mnist /

[36] B. Kwolek, "Face detection using convolutional neural networks and Gabor filters," in Lecture Notes in Computer Science, vol. 3696. 2005, p. 551.

[37]G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. Neural Computation, 18:1527–1554, 2006.

[38] S. Sukittanon, A. C. Surendran, J. C. Platt, and C. J. C. Burges, "Convolutional networks for speech detection," Interspeech, pp. 1077–1080, 2004.

[39] F.-J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional nets for generic object categorization," in Proc. Computer Vision and Pattern Recognition Conf. (CVPR'06), 2006.

[40] H. Lee, Y. Largman, P. Pham, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in Advances in Neural Information Processing Systems 22 (NIPS'09), 2009.

[41] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in Proc. Int. Conf. Computer Vision and Pattern Recognition, 2009.

[42] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in Proc. 25th Int. Conf. Machine Learning, 2008, pp. 1168–1175.

[43] K. A. DeJong, "Evolving intelligent agents: A 50 year quest," IEEE Comput. Intell. Mag., vol. 3, no. 1, pp. 12–17, 2008. [25] X. Yao and M. Islam, "Evolving artificial neural network ensembles," IEEE Comput. Intell. Mag., vol. 2, no. 1, pp. 31–42, 2008.

[44] I. Arel, D. C. Rose and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]," in IEEE Computational Intelligence Magazine, vol. 5, no. 4, pp. 13-18, Nov. 2010
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5605630&isnumber=5605610
[45] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In Proc. CVPR, 2014
[46] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In Proc. BMVC., 2013.
[47] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deep-Face: Closing the gap to human level http://www.darpa.mil/IPTO/solicit/baa/BAA0940_PIP.pdf
[48] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007
[49] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In Proc. CVPR, 2014. http://www.numenta.com
[50] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In Proc. ECCV, pages 566–579, 2012.
[51] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and
robust. CoRR, abs/1412.1265, 2014.
[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and
A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014. performance in face verification. In Proc. CVPR, 2014.
[53] http://cvlab.cse.msu.edu/project-dr-gan.html
[54]L. Tran, X. Yin and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1283-1292
[55]M. Coşkun, A. Uçar, Ö. Yildirim and Y. Demir, "Face recognition based on convolutional neural network," 2017 International Conference on Modern Electrical and Energy Systems (MEES), Kremenchuk, 2017, pp. 376-379.
[56] A. Uçar, Y. Demir, and C. Guzelis, “Object Recognition and Detection with Deep Learning for Autonomous Driving Applications,” Simulation,pp. 1-11, 2017.

[57] T. Bianchi and A. Piva, “Image forgery localization via block-grained analysis of JPEG artifacts,” IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 1003–1017, 2012.

[58] I. Amerini, T. Uricchio, L. Ballan, and R. Caldelli, “Localization of jpeg double compression through multi-domain convolutional neural networks,” Proc. of IEEE CVPR Workshop on Media Forensics, 2017.

[59] I. Amerini , T.uricchio , R.Caldelli , “Tracing images back to their social network of origin: a CNN-based approach” , National Inter-University Consortium for Telecommunications (CNIT), Parma, Italy

[60] H. Zhang, Z. Qu, L. Yuan and G. Li, "A face recognition method based on LBP feature for CNN," 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, 2017, pp. 544-547.

[61]. A. Rozantsev, V. Lepetit, and P. Fua. “On rendering synthetic images for training an object detector,” Computer Vision and Image Understanding, 2015.

[62] http://knowledge.esciencecenter.nl/content/LargeScaleComputerVision.pdf

[63]https://www.techleer.com/articles/444-capsule-networks-an-improvement to-convolutional networks /

[64]https://d9johdpvmzlgp.cloudfront.net/images/Blog/general-architecture-convnet.jpg

[65]Benjamin Graham « Fractional Max-Pooling [archive] », 18 décembre 2014

[66]A. Krizhevsky, I. Sutskever et G. E. Hinton, « ImageNet Classification with Deep Convolutional Neural Networks », Advances in neural Processing Systems de traitement, vol. 1, 2012, p. 1097–1105 lire en ligne archive du 16 février 2015

[67]https://optinum.co.za/whats-new-deep-learning

[68]http://neuralnetworksanddeeplearning.com/chap1.html#sigmoid_neurons

[69] http:// www.qingpingshan.com/bc/jsp/163284.html

[70] ImageNet. http://www.image-net.org

[71] Russakovsky, O., Deng, J., Su, H., et al. "ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision (IJCV). Vol 115, Issue 3, 2015, pp. 211–252

[72] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." Advances in neural information processing systems. 2012.

[73]Feature Extraction Using AlexNet,Help Matlab 2018

[74] http://deeplearning.net/deep-learning-research-groups-and-labs/

[75] Y. Lv, Z. Feng and C. Xu, "Facial expression recognition via deep learning," 2014 International Conference on Smart Computing, Hong Kong, 2014, pp. 303-308.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7043872&isnumber=7043829

[76] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, “Dexpression: Deep convolutional neural network for expression recognition,” arXiv preprint arXiv:1509.05371, 2015.